



# Analyzing Georeferenced Tweets to Explore Pre- and Post-COVID Urban Neighborhood Dynamics in New York City

## Master Thesis

for the attainment of the Master's degree "Master of Science",  
abbreviated "MSc"

submitted within the University Master Program for Further Education  
"Geographical Information Science & Systems – (UNIGIS MSc)"  
at the Department of Geoinformatics - Z\_GIS,  
Faculty of Digital and Analytical Sciences,  
University of Salzburg

submitted by

**B.Sc. Marjorie Mattes**

Supervisor:

Prof. Dr. Bernd Resch

Heidelberg, December 2025

## **Abstract**

This thesis investigates how geotagged Twitter data can be used to classify Tweets into predefined urban functional categories and analyze their temporal and spatial dynamics in New York City from 2018 to 2022. The study focuses on five key domains: Transportation, Retail Activity, Cultural/Social Activity, Healthcare and Work/Remote Work. To extract and classify content related to these categories, a methodological workflow combining Wikipedia-derived keyword filtering with BERTopic-based modeling was developed.

Temporal and spatial analyses of category-related activity reveal distinct patterns of intensifications and declines, particularly within categories like Transportation and Cultural/Social Activity. Deviations from long-term baseline shares illustrate localized disruptions and partial recoveries in these categories, while categories with sparse representation, such as Healthcare and Work/Remote Work, display fragmented patterns that limit interpretability.

The study demonstrates both the potential and the constraints of using geotagged social media activity as a complementary data source for understanding urban behavioral change. While representation biases and data sparsity remain challenges, the developed workflow offers a means to trace spatial and temporal shifts in selected aspects of urban life, particularly during large-scale events such as the COVID-19 pandemic.

# Table of Contents

List of Figures.....	3
List of Tables.....	4
1. Introduction.....	5
2. Related Work.....	6
2.1 Social Media for Urban Studies .....	6
2.2 Topic Identification from Short Texts.....	7
2.2.1 Traditional Probabilistic Models.....	7
2.2.2 Keyword-Based Topic Analysis .....	7
2.2.3 Adding Wikipedia as a Source for Semantic Context .....	8
2.2.4 Topic Modeling using Transformers.....	8
3. Methodology .....	9
3.1 Methodological Approach .....	9
3.2 Study Area and Data Sources .....	11
3.2.1 Twitter (now X).....	11
3.2.2 Study Area .....	11
3.3 Data Pre-Processing .....	12
3.3.1 Data Filtering .....	12
3.3.2 Text Normalization .....	12
3.4 Keyword Generation for Pre-filtering.....	12
3.4.1 Collection of Relevant Wikipedia Articles .....	13
3.4.2 SBERT-Based Document Clustering of Wikipedia Articles.....	14
3.4.3 Extracting Keywords per Cluster using TF-IDF.....	14
3.5 Topic Modeling of Tweets using BERTopic.....	14
4. Results .....	16
4.1 Results of Topic Modeling and Category Assignment .....	16
4.2 Spatial Distributions and Patterns of categorized Tweet Content throughout the years of 2018 to 2022 .....	20
4.2.1 Overview of all Categories.....	20
4.2.2 Transportation.....	22
4.2.3 Retail Activity.....	26
4.2.4 Cultural/Social Activity .....	29
5. Discussion .....	34
5.1 Discussion of Methodology .....	34
5.2 Discussion of Results .....	36

6. Conclusion ..... 39

Appendices ..... 40

References..... 55

**List of Figures**

Figure 1 Workflow for Preparation of Twitter Data, Keyword Extraction, Tweet Classification and Spatial Analysis ..... 10

Figure 2 Locations of NYC Boroughs and ZIP Code Tabulation Areas ..... 16

Figure 3 Development of geolocated Tweet Counts per Year: Original Dataset for New York Metropolitan Area (Blue) and filtered Dataset for New York City (Red) ..... 17

Figure 4 Category Shares among Category-Related Tweets in NYC..... 18

Figure 5 Log-Scale Counts for Categories in NYC ..... 19

Figure 6 Distributions of Category-Related Tweets in NYC from 2018 to 2022..... 21

Figure 7 Public Transportation Infrastructure NYC (NYC Office of Technology & Innovation, 2025a, NYC Office of Technology & Innovation, 2025b, NYC Office of Technology & Innovation, 2022b, NYC Office of Technology & Innovation, 2022a) ..... 22

Figure 8 Share of Transportation Tweets in 2018 (left) and Deviation of Share from Average of All Years (right) ..... 23

Figure 9 Share of Transportation Tweets in 2020 (left) and Deviation of Share from Average of All Years (right) ..... 24

Figure 10 Share of Transportation Tweets in 2022 (left) and Deviation of Share from Average of All Years (right) ..... 25

Figure 11 LISA Maps of Transportation Deviation from All-Year Average in NYC..... 26

Figure 12 Non-Vacant Storefronts and Major Commercial Centers in NYC (Department of Finance (DOF), 2025) ..... 27

Figure 13 Share of Retail-Activity Tweets in 2019 (left) and Deviation of Share from Average of All Years (right) ..... 28

Figure 14 Share of Retail-Activity Tweets in 2022 (left) and Deviation of Share from Average of All Years (right) ..... 29

Figure 15 Parks and Cultural-/Recreational Facilities in NYC (NYC Office of Technology & Innovation, 2024b, NYC Office of Technology & Innovation, 2024a)..... 30

Figure 16 Share of Cultural/Social Activity Tweets in 2018 (left) and Deviation of Share from Average of All Years (right)..... 31

Figure 17 Share of Cultural/Social Activity Tweets in 2020 (left) and Deviation of Share from Average of All Years (right)..... 31

Figure 18 Share of Cultural/Social Activity Tweets in 2021 (left) and Deviation of Share from Average of All Years (right)..... 32

Figure 19 LISA Maps of Cultural/Social Activity Deviation from All-Year Average in NYC ..... 33

**List of Tables**

Table 1: Selected Categories and the corresponding Wikipedia articles used for scraping linked articles ..... 13

Table 2 Precision Results per Category ..... 19

Table 3 Global Moran's I Results of Distribution of Deviation of Transportation Shares over All-Year Average..... 25

Table 4 Global Moran's I Results of Distribution of Deviation of Retail-Activity Shares over All-Year Average..... 29

Table 5 Global Moran's I Results of Distribution of Deviation of Cultural/Social-Activity Shares over All-Year Average ..... 32

## 1. Introduction

Cities, regardless of their size, geography or level of development, share organizational, social and economic characteristics and evolve as complex systems shaped by interdependent social, infrastructural and spatial processes (Bettencourt et al., 2013, Bettencourt, 2013, Batty et al., 2018).

However, as highlighted by Batty et al. (2022), cities are also sensitive to disruptions like the COVID-19 pandemic and adapt to the corresponding implications and restrictions, which alter the way of urban life and how people interact within cities.

These interferences affect different areas and trends of day-to-day life that range from remote work and commuting to the reconfiguration of retail and leisure spaces as well as digital communication (Florida et al., 2021, Batty et al., 2022, Glaeser, 2022, Bandarin et al., 2021, Rao et al., 2022, Megahed and Abdel-Kader, 2022).

The concept of the Post-COVID City (Batty et al., 2022) provides a useful framework for interpreting these transformations and identifying key functional domains that were most affected during and after the pandemic like mobility, healthcare, retail, work and cultural and social activity. Leveraging this conceptual foundation, this thesis uses these domains as thematic anchors to explore how urban social media activity evolved in New York City between 2018 and 2022 regarding these functions.

Social media platforms such as X (formerly Twitter) can provide a bottom-up view of public online activity, recording perceptions and everyday practices in near real-time (Goodchild, 2007, Crooks et al., 2015). This thesis applies a combined keyword filtering and BERTopic-based approach to identify Tweets related to the main functional domains of urban life and to analyze how these discussions manifest spatially and temporally.

This work first develops and evaluates a methodological workflow for filtering and classifying geolocated Twitter data to capture category-related content within the urban life domains Transportation, Retail Activity, Cultural/Social Activity, Healthcare and Work/Remote Work.

Building on this foundation, it then examines how category-related activity changed in New York City between pre-pandemic, pandemic, and post-pandemic periods and to what extent deviations from long-term baseline shares reveal spatially clustered disruptions or recoveries linked to these shifts in urban social media activity.

Accordingly, this thesis is guided by the following research questions:

**(RQ 1)** *How can geotagged Twitter posts be systematically filtered and classified into the urban life categories of Transportation, Retail Activity, Cultural/Social Activity, Healthcare and Work/Remote Work?*

**(RQ 2)** *How did categorical Tweet activity in New York City change between pre-pandemic, pandemic and post-pandemic periods?*

**(RQ 3)** *How do the spatial distributions of categorical Tweet activity change across pre-pandemic, pandemic, and post-pandemic periods, and where do localized intensifications or declines become visible?*

## **2. Related Work**

### **2.1 Social Media for Urban Studies**

Crowd-contributed data from social media, open source or volunteered datasets provide first-hand insights into functional developments in urban space in a fine spatial, temporal and social resolution (Goodchild, 2007, Crooks et al., 2015). The use of people's digital footprints has gradually emerged in different disciplines over the past few years and has quickly become a main source for data-driven analysis and research (Yang and Liu, 2022, Miller and Goodchild, 2015, Goodchild, 2007, Niu and Silva, 2020).

Since there has been significant amount of research on crowdsourced data, a few notable studies from applications such as human mobility and activity pattern analysis, sentiment and perception analysis and urban functional use detection (Niu and Silva, 2020) will be described.

The review work of Niu and Silva (2020) and Yang and Liu (2022) shows that the forms of available crowdsourced data and the ways it can be processed and utilized are manifold. Liu et al. (2014) for example use social media data to derive trajectory patterns of trips to display interaction strength between Chinese cities, human mobility patterns and spatially embedded networks.

Guo et al. (2016a) investigated a unigram-based sentiment analysis with geotagged Tweets for different social groups for the Greater London area and found that happiness is linked to urban features and socioeconomic parameters such as number of jobs, children and transportation possibilities. Kovacs-Györi et al. (2018) performed analyses on Tweets to gain new insights about visitors and spatial, temporal and affective patterns of park visits in London and found that positive sentiments are more frequent in parks than in other urban areas.

Steiger et al. (2016) proposed a self-organizing map to visualize spatiotemporal clusters of human activity that also consider semantic similarities of Twitter posts via the Latent Dirichlet allocation model (LDA) (Blei et al., 2003) to uncover further mobility patterns.

Researchers like De Sabbata et al. (2023) used advanced Natural Language Processing (NLP) methods to distinguish everyday life social media content from event-related content in the city of Leicester. Authors like Huang et al. (2022) applied this line of approach within the COVID-19 pandemic context by classifying and analyzing Tweet content in regards to urban parks.

This thesis aims to build on this work by integrating multiple analytical layers using semantic models to understand complex social media activity around selected urban categories.

## **2.2 Topic Identification from Short Texts**

### *2.2.1 Traditional Probabilistic Models*

The exploration of unstructured textual information made use of different text mining methods (Steiger et al., 2016) where topic modeling has proven to be one of the most popular ones according to Karami et al. (2020) as it aims to disclose the hidden semantic structure of the text (Karami et al., 2020).

A foundational approach within this field is probabilistic topic modeling, exemplified by Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA operates on a "bag-of-words" assumption, treating documents as a mixture of topics and topics as a distribution of words. It has been widely applied in urban studies to identify topical patterns related to tourism, disasters, and daily activity from large text corpora like online reviews and social media data (Guo et al., 2016b, Resch et al., 2018, Lansley and Longley, 2016).

Regardless of the popularity, researchers criticize that LDA tends to neglect co-occurrence relations within a document and that noisy and sparse datasets, like Tweets, are unsuitable for LDA (Egger and Yu, 2022, Chen et al., 2019).

### *2.2.2 Keyword-Based Topic Analysis*

Many approaches however also resort to the use of direct keyword filtering to retrieve data that contains relevant information. Authors like de Albuquerque et al. (2015), Guan and Chen (2014) adapted this approach to compare the spatial and temporal distribution of Tweets in regards to natural disaster events.

To bridge possible limitations of manual keyword lists, further approaches aim to generate keywords automatically or expand these semantically using external knowledge sources. The following section will discuss how Wikipedia, in particular, has been leveraged for this purpose.

### *2.2.3 Adding Wikipedia as a Source for Semantic Context*

According to Biuk-Aghai and Ng (2014), the use of Wikipedia data for the creation of ontologies, keyword extraction and mapping topics has increased alongside its growth in scope and quality content and is now one of the largest digital encyclopedias worldwide that has a broad knowledge coverage about different concepts (Rath and Chow, 2022, Wu et al., 2017, Gabrilovich and Markovitch, 2007).

The approaches on salvaging Wikipedia's data for semantic context are manifold. Biuk-Aghai and Ng (2014) for example, developed a method for analyzing and automatically classifying publications by using the Wikipedia category hierarchy and matching them to relevant keywords extracted from corresponding Wikipedia articles using TF-IDF (Ramos, 2003). TF-IDF (Term Frequency-Inverse Document Frequency), is a measure to evaluate the importance of a word within a collection of documents (Salton and Buckley, 1988) and remains widely used in modern topic modeling as described by Egger and Yu (2022).

Similarly, Wu et al. (2017) aim to address the issue of semantic mismatch problems by proposing a technique named "Wikipedia Matching", which uses a large number of concepts from Wikipedia to construct a reference space containing concept vectors holding the semantic information.

In more recent research, modern methods employ transformer-based models for information extraction such as the work of Rath and Chow (2022), who use Sentence-BERT (SBERT) to extract useful information from relevant keylines within Wikipedia articles in the context of city typology prediction.

The described works illustrate how Wikipedia can be used as semantic resource for different tasks, underlining its potential for keyword extraction in this thesis.

### *2.2.4 Topic Modeling using Transformers*

The advent of the Transformer architecture (Vaswani et al., 2017) and pre-trained language models like BERT (Devlin et al., 2018) revolutionized NLP by providing deep contextual understanding of text. As described by Patwardhan et al. (2023) this technology has paved the way for a wide range of downstream tasks like transformer-based topic modeling methods.

Building on this technological foundation, BERTopic has emerged as a prominent topic modeling technique (Grootendorst, 2022). In contrast to other models like LDA, Grootendorst (2022), Egger

and Yu (2022) not only emphasize BERTopic's embedding approach as a feature that reduces the pre-processing necessity of the data, but also its capability to uncover contextually rich topics, while also providing clear distinctions between the identified topics.

Its effectiveness in analyzing complex, user-generated short texts has led to its use in recent studies as well as in this thesis. For example, Jiang et al. (2023) applied BERTopic to display and understand content about urban shrinkage through local newspapers in the city of Detroit. Researchers like Xu et al. (2022) used BERTopic to reveal changes in Twitter content on masks within different political user groups for the year 2020. In a similar vein, De Sabbata et al. (2023) use BERTopic to reveal everyday life related topics in the city of Leicester. G. Almatar et al. (2020) assess the primary topics of interest of Tweets in Kuwait for a week and its spatiotemporal distributions.

### **3. Methodology**

#### **3.1 Methodological Approach**

This thesis uses a multi-stage NLP pipeline (Figure 1) consisting of four main components. The first one focuses on the preparation of geolocated Twitter data, including filtering, normalization and removal of organizational accounts. The second component encompasses keyword extraction using Wikipedia-based article scraping and SBERT clustering as well as TF-IDF for keyword selection. The third component includes the classification of Tweets through keyword-based filtering and BERTopic-based topic modeling, followed by manual mapping of BERTopic topics to predefined urban functional categories. The final component consists of spatial and temporal evaluation of the category-related Tweet activity in New York City between 2018 and 2022.

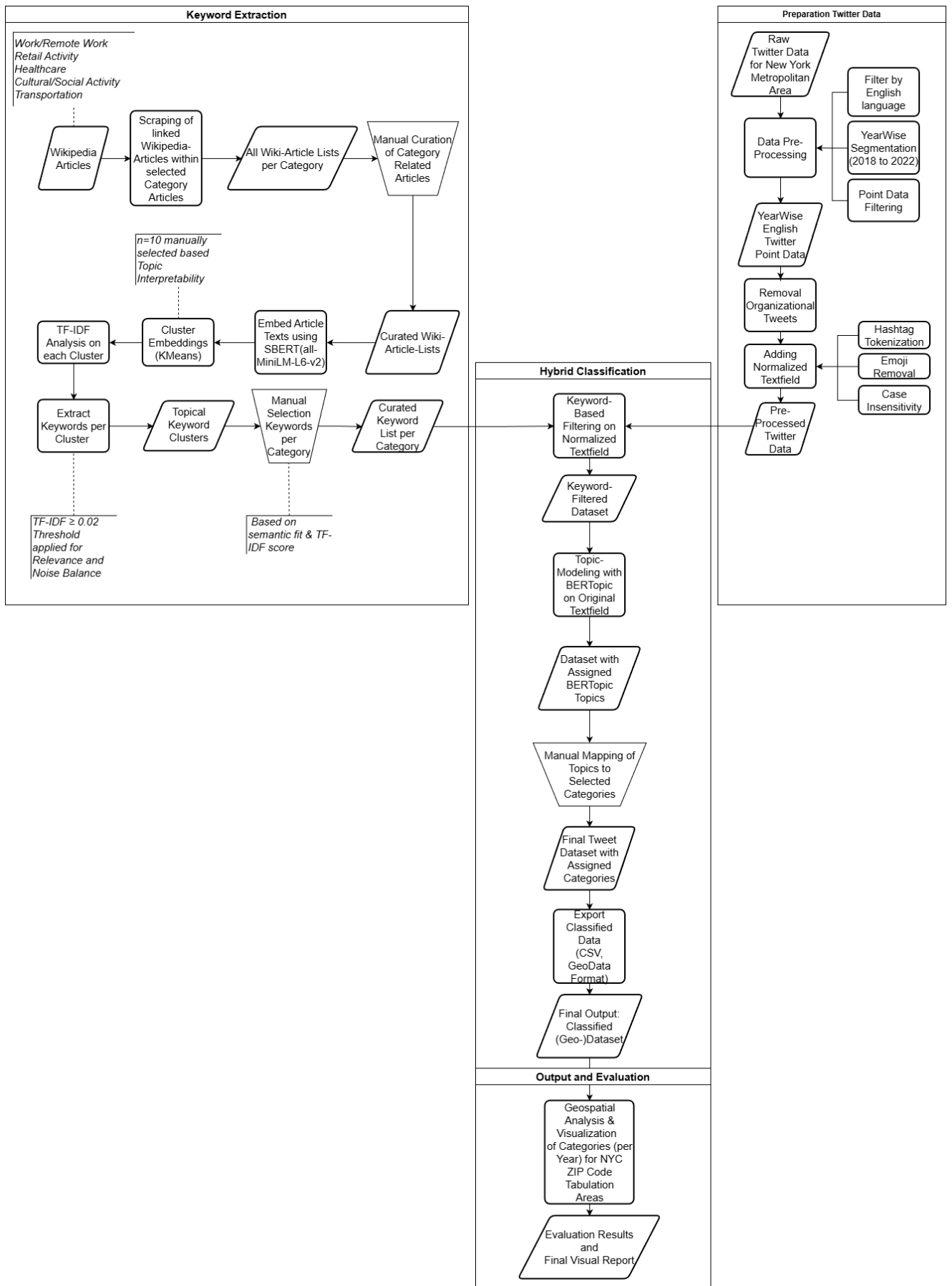


Figure 1 Workflow for Preparation of Twitter Data, Keyword Extraction, Tweet Classification and Spatial Analysis

## 3.2 Study Area and Data Sources

### 3.2.1 Twitter (now X)

The dataset consists of public posts made on the former Twitter platform (now X), a microblogging service where users share short messages of up to 280 characters which can contain text, emojis, URLs and other media (Karami et al., 2020, Asgari-Chenaghlu et al., 2021, Roberts et al., 2018). The data used is composed of public tweets posted between January 1, 2018 and December 31, 2022 for the City of New York. It has been collected via application programming interfaces (API) using a proprietary Java-based crawler and provided by the University of Salzburg.

The metadata of a single tweet usually contains additional information next to the textual content such as a timestamp, the display name of the user and their profile location text and, if shared, the GPS geotag (Nguyen et al., 2022, Asgari-Chenaghlu et al., 2021). These geotags are the most obvious source of location information, as they are provided in form of longitude and latitude coordinates that represent a precise geographical location without further processing needed (Nguyen et al., 2022, Asgari-Chenaghlu et al., 2021).

### 3.2.2 Study Area

With an estimated population of 8,804,190 distributed over 300.40 square miles (778.03 km<sup>2</sup>), New York City is one of the most densely populated major cities in the United States (U.S. Census Bureau), where global crises like the COVID-19 pandemic can have magnified effects (Ignaccolo et al., 2024).

The period of 2018 throughout 2022 was marked by several pivotal events that shaped urban life and public discourse in New York City. The initial phase from 2018 up to the start of 2020 provides a pre-pandemic baseline in regards to the data that will be examined within this thesis.

The first cases of COVID infections reached the United States in January 2020 and around the time the World Health Organization (WHO) declared a global pandemic on March 11, 2020, 355 cases were confirmed in New York City (NYC Department of Health and Mental Hygiene (DOHMH), Ignaccolo et al., 2024). Lockdown and social distancing measures were applied and impacted different sectors like gastronomy, working life, education, infrastructure and social life greatly (U.S. Centers for Disease Control and Prevention, Ignaccolo et al., 2024, Cuomo, March 20, 2020).

After the widespread administration of COVID vaccines by the end of 2020 and the steady distribution among the public, restrictions were slowly withdrawn at the start of 2021 (Cuomo, December 21, 2020). By June 2021, the testing positivity rate of the city reached its lowest since the start of the pandemic and the entirety of the New York State was reopened by Governor Cuomo (Cuomo, June 15, 2021).

### **3.3 Data Pre-Processing**

#### *3.3.1 Data Filtering*

As it is the aim to analyze geotagged Tweets (Goodchild, 2007, Ignaccolo et al., 2024), only Tweets filed with a point geometry and corresponding latitude/longitude coordinates were used in the final dataset, which resulted in a total of 15,005,830 Tweets for the New York metropolitan area for the years 2018 throughout 2022.

As the chosen NLP models are optimized for the English language (Conneau et al., 2019), the data was filtered down to only retain Tweets that have the corresponding label filed in the language field.

Furthermore, the focus lay on narrowing down the content of individual X users rather than organizations, which is why Tweets of potential institutional or automated accounts were removed making use of the profile metadata (Ferrara et al., 2016).

To achieve this, a heuristic-based filtering method described in Nagarkar et al. (2020) was adapted. In their approach, Nagarkar et al. (2020) removed business accounts by first identifying indicative keywords such as “realtor”, “developer”, or “group” and then filtering out accounts whose usernames matched a manually curated list of such business-related terms.

Following this, this thesis adopted a similar heuristic approach by creating a list of keywords (“news”, “radio”, “journal”, “jobs”, “tmj\_”, “tmj”, “511”, “traffic”) and filtering out those Tweets where usernames containing those terms were classified as organizational and excluded from the final dataset to ensure the resulting corpus is more representative of authentic individual expression. These keywords were selected because of their consistent appearance in institutional or automated accounts in several inspection rounds.

#### *3.3.2 Text Normalization*

To ensure that content of hashtags as well as word combinations with emojis were taken into account during the upcoming keyword filtering process, a secondary version of the text field was created. This text field contained the normalized version of the text, where hashtags were tokenized using the Python package Wordninja (Keredson, (n. d.)), emojis removed with the Emoji Python package (Carpedm, (n. d.)) and case-insensitivity applied.

### **3.4 Keyword Generation for Pre-filtering**

While powerful, a purely unsupervised topic modeling approach like BERTopic may not produce topics that align well with the predefined categories. To support the model and ensure the relevance of the final corpus, a pre-filtering step based on the extraction of relevant keywords was employed.

### 3.4.1 Collection of Relevant Wikipedia Articles

The approach is conceptually based on previous research like Biuk-Aghai and Ng (2014), who established a methodology for topic representation by extracting weighted keywords from relevant Wikipedia articles using TF-IDF.

To maintain computational efficiency and ensure thematic relevance (Wu et al., 2017), a small set of Wikipedia articles were selected as entry points. The selected seed articles correspond to the five functional categories derived from the Post-COVID City framework by Batty et al. (2022).

From these, all linked articles embedded within the HTML structure were scraped using the BeautifulSoup Python package (Richardson, 2024). The starting articles are listed in Table 1.

Category	Used Category-Related Main Articles for Scraping
Healthcare	Healthcare COVID-19
Work/Remote Work	Work_(human_activity) Job Remote_work
Retail Activity	Retail
Cultural/Social Activity	Culture Society Entertainment List_of_buildings,_sites,_and_monuments_in_New_York_City List_of_museums_and_cultural_institutions_in_New_York_City List_of_parks_in_New_York_City
Transportation	Transportation Traffic

**Table 1: Selected Categories and the corresponding Wikipedia articles used for scraping linked articles**

A final manual curation of the article lists was conducted to ensure the relevance of the corpus for the keyword extraction by removing articles that were not related to the functional categories, such as country names.

The extracted and curated Wikipedia articles are filed under Appendix 9 to Appendix 13.

The names of tourist attractions from the lists of buildings, museums and parks in New York were immediately filed as key terms for the upcoming keyword filtering procedure.

### *3.4.2 SBERT-Based Document Clustering of Wikipedia Articles*

As a next step, the approach made use of a well-performing pre-trained Sentence-BERT (SBERT) model all-MiniLM-L6-v2 (Reimers and Gurevych, 2019, Galli et al., 2024) to embed the Wikipedia article corpora for each list of articles per selected category and to identify underlying semantic structures and meaning.

Following this, a K-Means clustering algorithm was applied to these embeddings to divide the articles into 10 distinct, semantically coherent clusters, where each cluster represents a unique theme. Lower cluster numbers produce overly broad themes, while higher numbers often fragmented semantically coherent groups. Setting the amount to 10 clusters provided interpretable clustering themes throughout all article sets per category.

### *3.4.3 Extracting Keywords per Cluster using TF-IDF*

After partitioning the documents, a TF-IDF vectorizer was applied within each cluster to identify terms that are most relevant to the theme the cluster represents. A minimum TF-IDF score of 0.02 was chosen because lower thresholds introduced large amounts of generic or unspecific keywords, while higher thresholds tended to remove domain-specific ones. A threshold of 0.02 yielded more fitting keyword sets across all clusters, requiring only minimal manual curation in the subsequent step.

The final sets of keywords per category were then created by manually selecting the most representative, high-scoring terms from the 10 thematic clusters and filed in a final nested keyword list.

The Tweets were then filtered according to the existence of matching keywords within their normalized text field. Tweets that did not contain any of the keywords provided in the keyword list weren't used in the upcoming topic modeling step. However, strict keyword matching can result in ambiguity, particularly when Tweets contain terms associated with multiple functional categories. Therefore, the subsequent topic modeling step was employed to resolve potential overlaps and ensure a higher classification precision.

## **3.5 Topic Modeling of Tweets using BERTopic**

BERTopic was selected over traditional models like LDA due to its ability to leverage contextual word embeddings, enabling a more accurate representation of semantic meaning and reducing the need for extensive pre-processing, such as lemmatization and stopword removal (Grootendorst, 2022). In this study, BERTopic was applied in its default configuration on the original English Tweet text fields.

According to its developer, the model operates in three main stages (Grootendorst, 2022). First, the conversion of every document to its numerical embedding representation takes place using a pre-trained language model (Grootendorst, 2022). Since these embeddings are displayed in high dimensionality space, dimensionality is subsequently reduced with UMAP (Uniform Manifold Approximation and Projection)(McInnes et al., 2018) to improve efficiency.

Afterwards, the reduced embeddings are grouped into clusters by using HDBSCAN (Hierarchical Density Based Clustering of Applications with Noise), which identifies groups of semantically similar Tweets while designating unassignable or noisy content as outliers under topic “-1” (Campello et al., 2015, Campello et al., 2013).

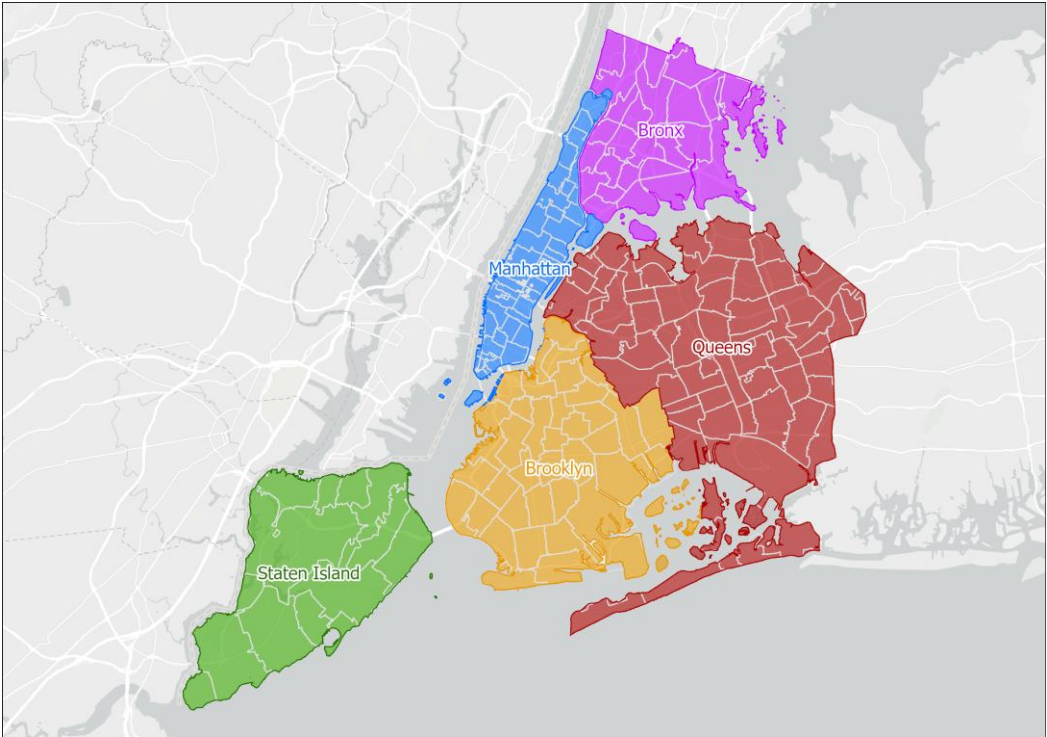
Finally, the topics are extracted using a class-based variation of TF-IDF (Term Frequency-Inverse Document Frequency) procedure (Grootendorst, 2022, Joachims, 1997).

To speed up the dimensionality reduction and clustering step, the cuML (Rapidsai, (n. d.)) version of UMAP and HDBSCAN was leveraged through GPU acceleration. A random seed state with the value 42 was set for reproducibility.

## 4. Results

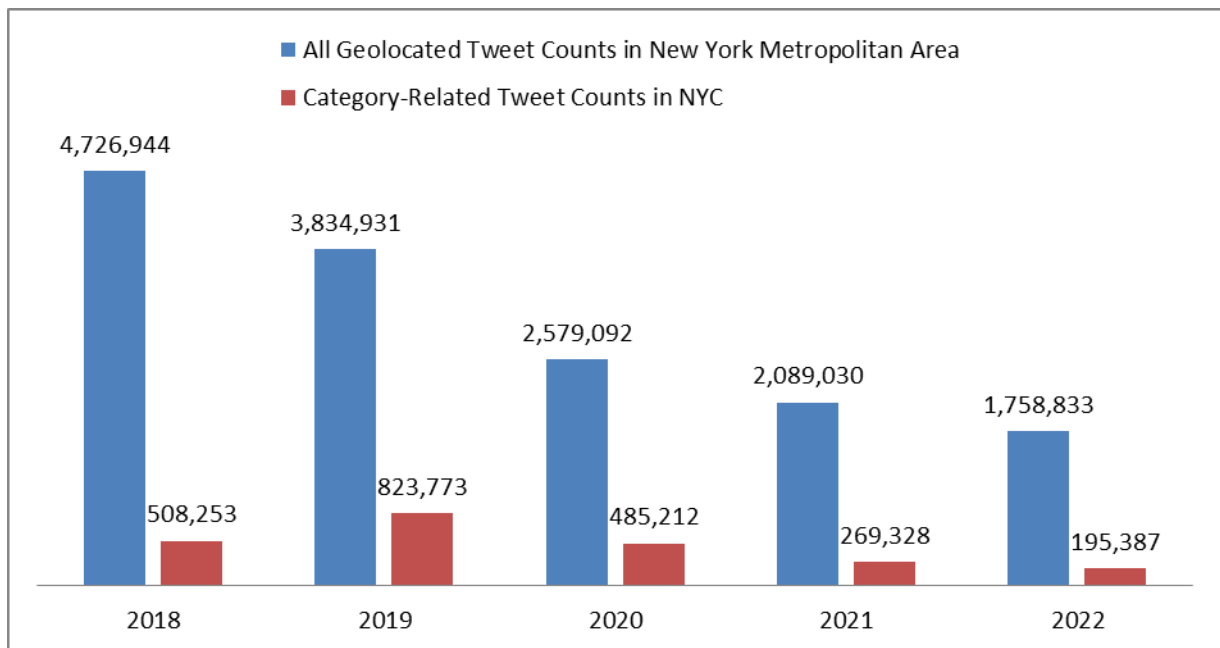
### 4.1 Results of Topic Modeling and Category Assignment

Following the Topic Modeling procedure 24% of Tweets were retained from the 15,005,830 geolocated Tweets in the New York Metropolitan Area. Focusing on an area of interest that encompasses the New York City boroughs Brooklyn, Bronx, Manhattan, Queens and Staten Island and their respective ZIP-Code-Tabulation-Areas (Figure 2), the final set of Tweets contains a total of 2,281,953 Tweets that represent the category-related Tweet counts for the years 2018 through 2022 and represent about 15% of the original Dataset.



**Figure 2** Locations of NYC Boroughs and ZIP Code Tabulation Areas

The number of category-relevant Tweets reached its peak in 2019, followed by a noticeable decline of Tweets in the following years. The general decline of geolocated Tweets compares to the general development of the New York Metropolitan Area (Figure 3).

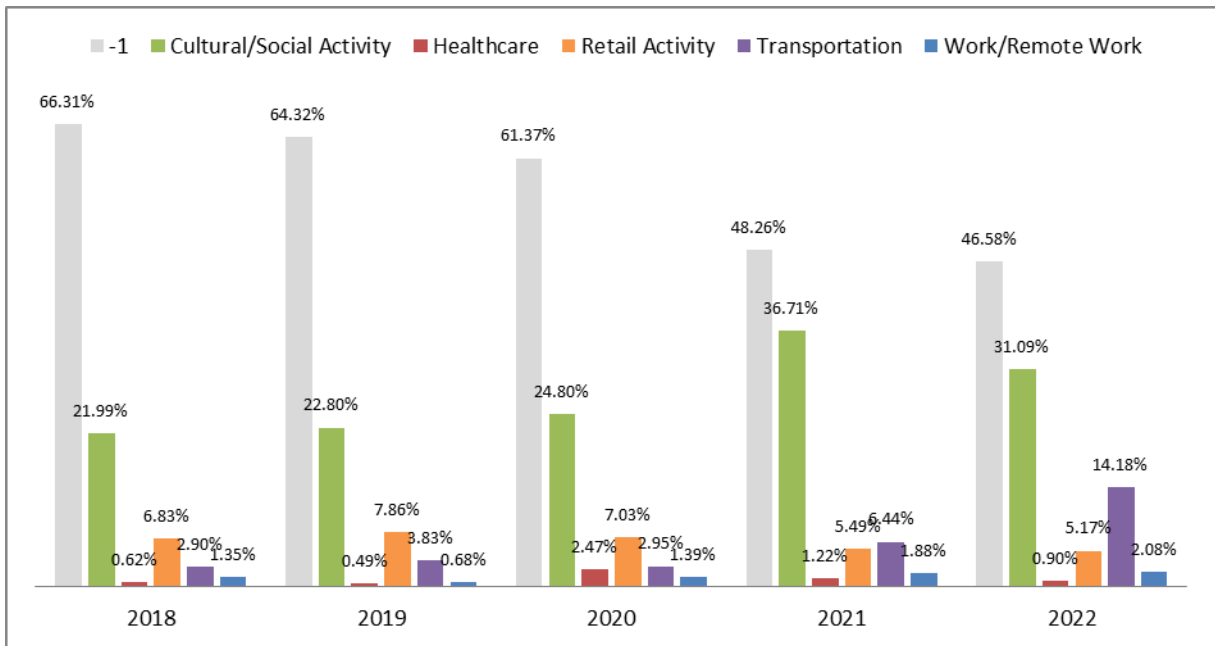


**Figure 3 Development of geolocated Tweet Counts per Year: Original Dataset for New York Metropolitan Area (Blue) and filtered Dataset for New York City (Red)**

Out of these Tweets, BERTopic assigned 39.28% Tweets to distinct topics, while 60.72% were recognized as outliers. After mapping the discovered BERTopic topics to the urban categories, Cultural/Social Activity emerged as the largest share representing 21.99% to 36.71% of category-related Tweets, peaking in the year 2021 in that regard (Figure 4).

Retail Activity followed with its highest share of 7.86% in 2019 and a continuous decline down to 5.17% in 2022. Transportation shows a strong share increase from 2.90% in 2018 to 14.18% in 2022.

The two remaining categories cover smaller share ranges, with Healthcare peaking in 2020 with a share of 2.5% and otherwise covering a range of 0.48 to 1.23%. Tweets classified as Work/Remote Work have their lowest share in 2019 with 0.67% and increasing up to 2.08% in 2022 (Figure 4).



**Figure 4 Category Shares among Category-Related Tweets in NYC**

While relative shares highlight the thematic composition of the Tweet content, Figure 5 presents the absolute category counts on a logarithmic scale to visualize the variations of category intensity across years and illustrating smaller categories like Work/Remote Work or Healthcare more clearly.

At this point, the latter shows a peak in Tweet counts in the year 2020 while larger categories like Cultural/Social- and Retail Activity show a decline from that year on. The strong increase of Transportation-related Tweets is underlined in this illustration as well.



**Figure 5 Log-Scale Counts for Categories in NYC**

To assess the quality of the category assignment, a precision evaluation was performed based on a random sample of 200 Tweets per category. Precision values indicate the share of correctly classified Tweets after manual validation. As shown in Table 2, all categories achieved precision scores above 90%, suggesting that the combined keyword filtering and BERTopic-based workflow provides reliable classification results across the selected themes.

Category	Precision
Healthcare	91.00%
Transportation	95.50%
Retail Activity	93.00%
Work/Remote Work	92.50%
Cultural/Social Activity	94.50%

**Table 2 Precision Results per Category**

## **4.2 Spatial Distributions and Patterns of categorized Tweet Content throughout the years of 2018 to 2022**

### *4.2.1 Overview of all Categories*

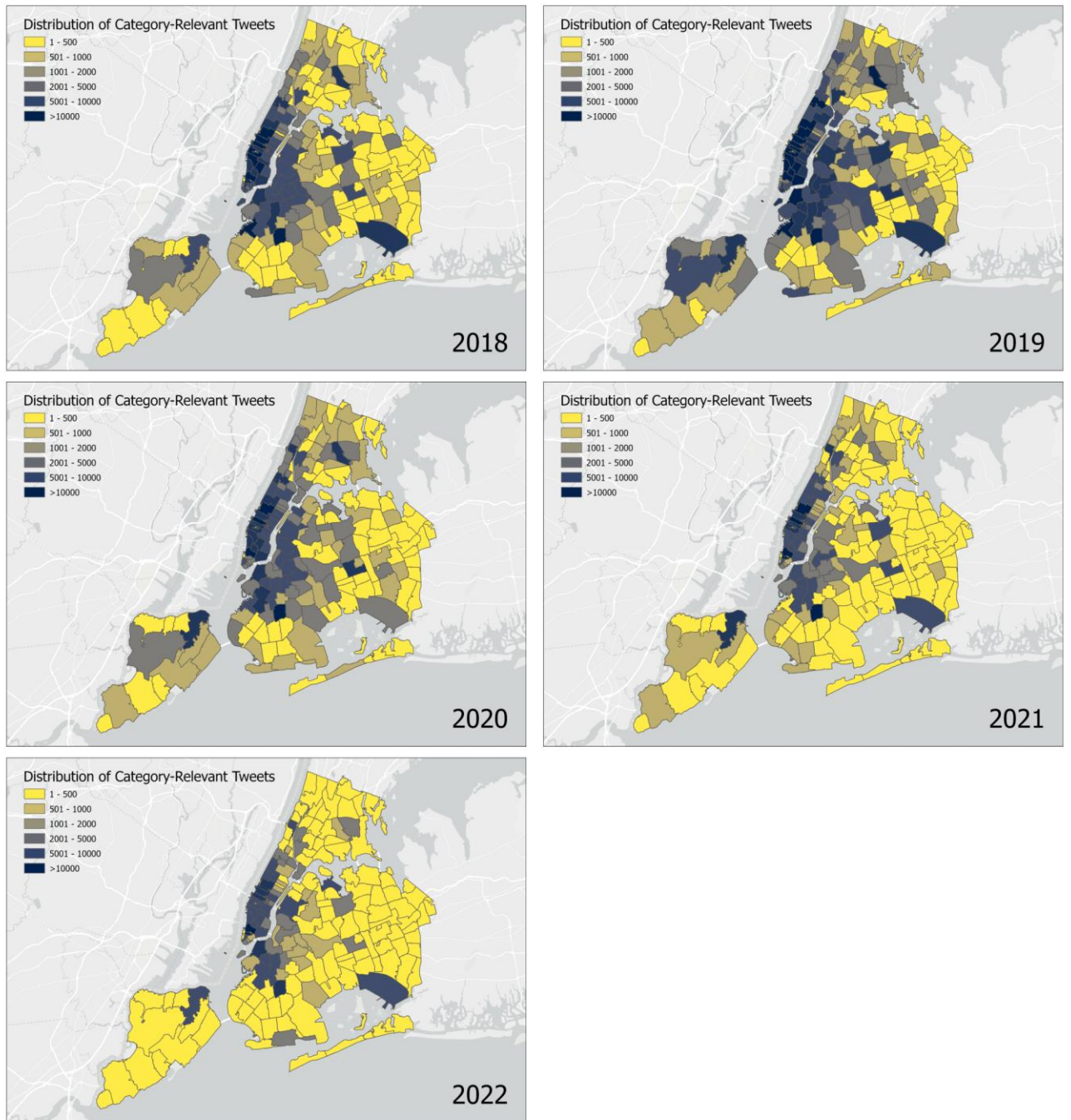
As seen in Figure 6, the spatial distribution of Tweets classified towards a category shifts notably between 2018 and 2022. In the pre-pandemic years of 2018 and 2019, ZIP codes within Manhattan and adjacent areas of the Bronx, Brooklyn or Queens dominate the picture with many areas showing at least 5,000 and in some cases even more than 10,000 category-related Tweets per year.

The John F. Kennedy International Airport (JFK Airport) in southern Queens also emerges as a prominent location, while Staten Island shows higher counts close to the Upper New York Bay.

In 2020, volumes begin to decline, especially in areas that showed high volume counts in the previous year. At this point, several ZIP Codes in Manhattan still exceed 5,000 Tweets but in the Bronx, Brooklyn, Queens, and Staten Island such counts become rare, with only a few isolated areas surpassing this threshold.

This contraction of category-related Tweet activity becomes more visible in 2021, where a few ZIP codes in Manhattan and nearby parts of the Bronx, Brooklyn and Queens exceed 1,000 Tweets, while larger parts of the city fall below. Exceptions remain visible in the area around JFK Airport, New York Upper Bay and the area around Prospect Park in Brooklyn.

The overall decline continues in 2022, with Tweet counts thinning out across nearly all boroughs in New York City. Manhattan and its immediate surroundings still account for the highest Tweet count volumes, but many ZIP codes in the Bronx, Brooklyn, Queens and Staten Island do not even exceed 500 category-related Tweets.



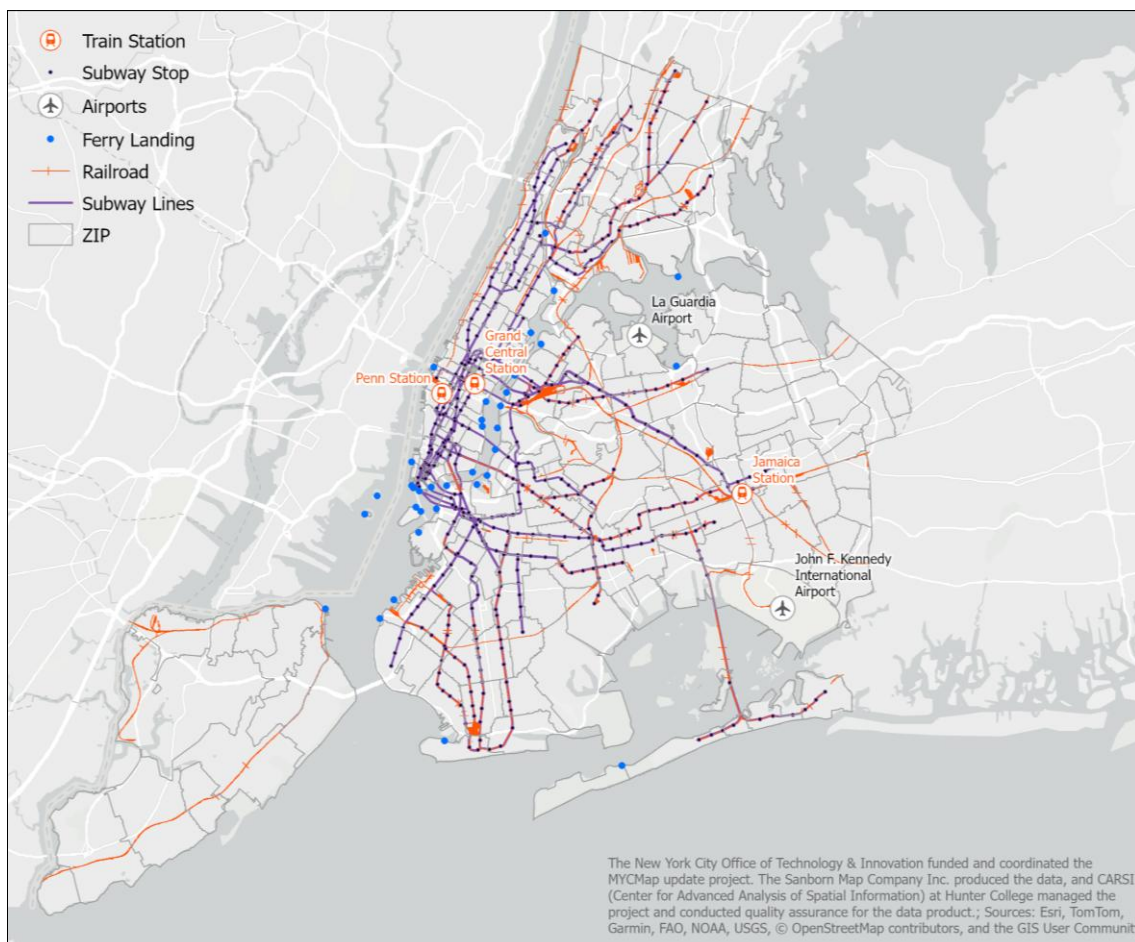
**Figure 6 Distributions of Category-Related Tweets in NYC from 2018 to 2022**

While Figure 4 and Figure 5 have illustrated the temporal development of category shares and counts in New York City, this work aims to analyze how these dynamics manifest spatially. Of the five functional categories, Healthcare and Work/Remote Work account for comparatively small volumes and show less distinct clustering. Therefore, the following spatial analysis focuses on Transportation, Retail Activity and Cultural/Social Activity, which provide a stronger data foundation and reveal clearer spatial dynamics over time. Maps and tables related to all topics can be found in the appendices.

#### 4.2.2 Transportation

Transportation-related Tweets encompass various kinds of content like human activity in public transit (“Sometimes it takes a while for the bus to show up in Kingsbridge. Right, mta? ??? @ Kingsbridge, Bronx”; JFK was a ghost town last week when I picked up daughter from SLC. #GladShesHome #AllGatheredIn #BeSafe @ JFK International Terminal 4”), but also traffic reports alongside location shares (“I’m at MTA Subway - Sutphin Blvd (F) - @nyctsubway in Jamaica, NY”; “Toyota RAV4 uber driver [plate redacted] blocked the bike lane near 1102 Bedford Ave on January 6 and has been reported to #nyctaxi. This is in Brooklyn Community Board 03 & #NYPD79. #VisionZero #BlockedBikeNYC”).

To contextualize the spatial patterns of transportation-related Tweets, Figure 7 provides an overview of New York City’s major public transportation infrastructure, including subway lines, train stations, ferry terminals and airports.



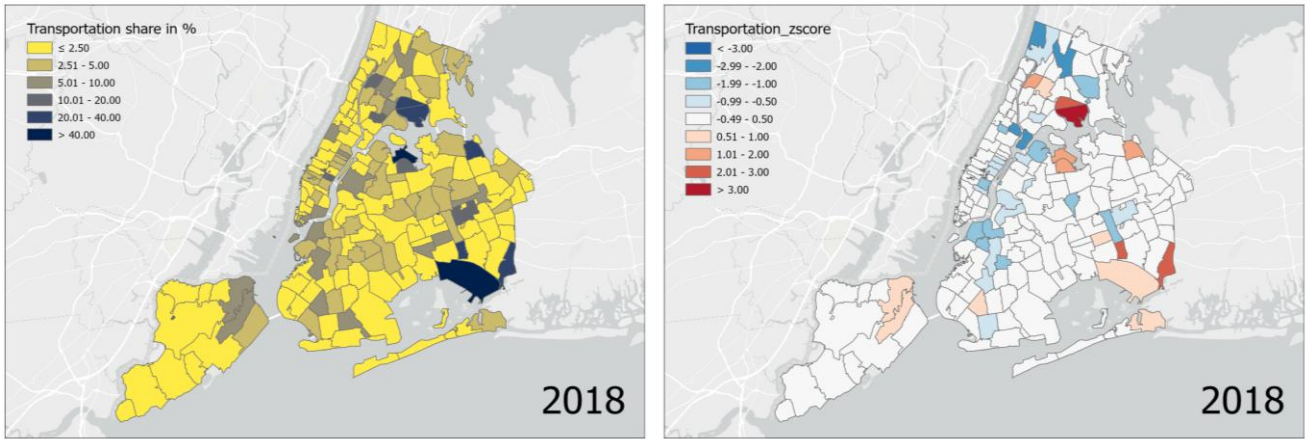
**Figure 7 Public Transportation Infrastructure NYC (NYC Office of Technology & Innovation, 2025a, NYC Office of Technology & Innovation, 2025b, NYC Office of Technology & Innovation, 2022b, NYC Office of Technology & Innovation, 2022a)**

In the pre-pandemic years 2018 and 2019, content about transportation seems widely scattered across the New York City boroughs, showing areas containing airports with high shares in Queens and Brooklyn and areas in the Bronx crossing the Henry Bruckner Expressway. At this point, the latter showed rather low Tweet counts in 2018, which suggests that the observed intensity might be more reflective of local Tweet scarcity rather than a sustained spatial cluster of transportation-related content.

Areas in proximity to subway or railway lines and their respective stations slightly stand out in Queens and Brooklyn while Staten Island shows many ZIP codes with very low shares of transportation-related Tweets.

This might be an indication of limited public transport options, as Staten Island is the only borough that does not have a connection to the New York City Subway System. Exceptions can be seen in areas that are close to the railroad or to the ferry station in the New York Upper Bay, connecting Staten Island with other boroughs.

Manhattan also shows quite low shares of transportation Tweets even though its public infrastructure is well connected while also harboring the Grand Central Terminal Station and the Penn Station, both being the two largest commuter rail hubs in New York City (Figure 8).



**Figure 8 Share of Transportation Tweets in 2018 (left) and Deviation of Share from Average of All Years (right)**

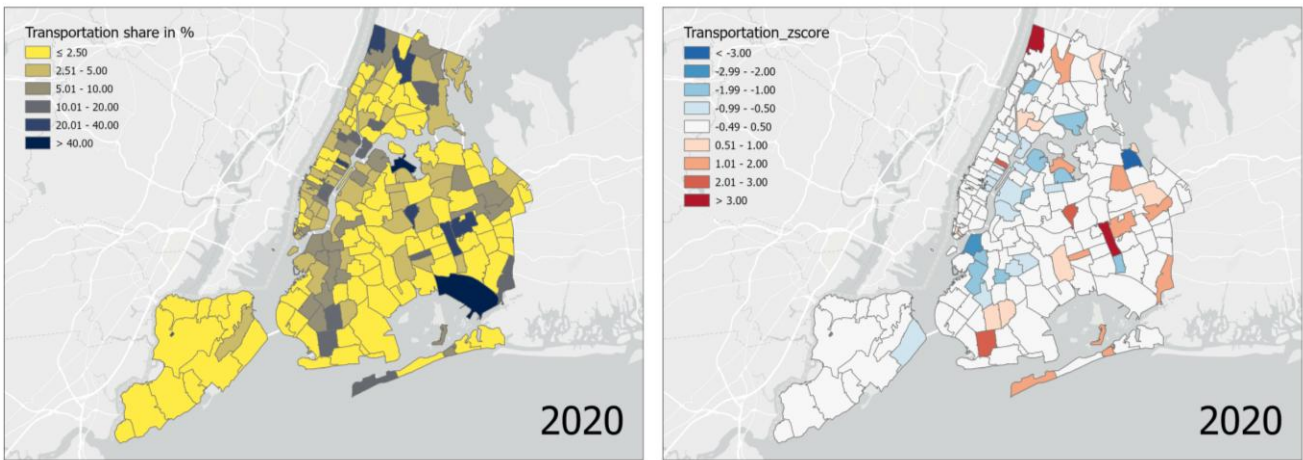
Even though the shares show where transportation-related Tweets are prominent in 2018, Figure 8 shows whether these values are above or below the long-term average in regards to all years. A map series visualizing the all-year average distributions per functional category is provided in Appendix 1.

The deviations highlight above-average values in the mentioned airport and expressway areas for 2018. It also becomes clear, that in Manhattan as well as central and northern Brooklyn,

transportation-related Tweets might be below the all-year average in pre-pandemic time, while Staten Island shows above-average activity close to public transportation hubs.

In 2019, the overall distribution remains comparable to 2018, with local exceptions such as above-average activity around Jamaica Station, but no sustained change in broader patterns.

At the onset of the COVID-19 pandemic in March 2020, the spatial distribution of transportation-related Tweets is seemingly reduced (Figure 9), which might indicate impacts of stay-at-home orders and cuts in public transport services leading to a great decline of commuter travel (Cuomo, March 20, 2020).

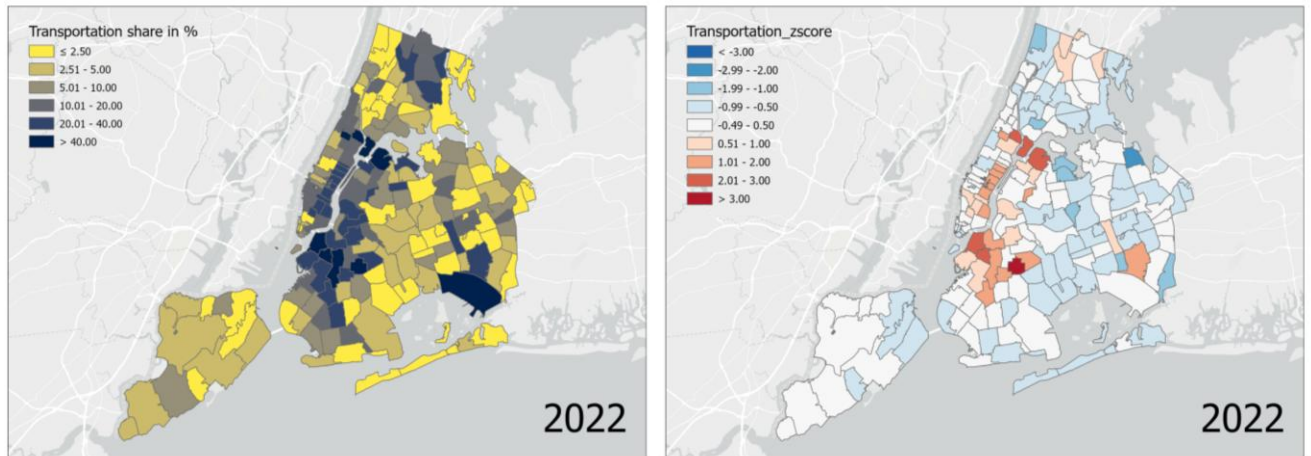


**Figure 9 Share of Transportation Tweets in 2020 (left) and Deviation of Share from Average of All Years (right)**

However, shares remain high around the main airports and Jamaica Station in Queens, while in central Brooklyn activity aligns more clearly with subway and railroad corridors. Staten Island continues to show predominantly low shares of transportation Tweets.

In contrast, the deviation map illustrates a patchier distribution of above- and below-average values across boroughs. Jamaica Station stands out with above-average activity, whereas JFK Airport, despite high overall shares, registers only average values.

By 2021, the first signs of a spatial shift appear that might indicate the restoration of full public transit in New York City (Cuomo, February 15, 2021). While Staten Island shows an increase in shares that remain in an all-year average range, Manhattan, Brooklyn and Queens, start to show above-average deviations as well as higher shares in public transport-linked corridors. This tendency intensifies in 2022, where above-average deviations and higher shares concentrate more clearly around subway as well as rail hubs and their accumulations are more distinguishable from widespread average or below-average values (Figure 10).



**Figure 10 Share of Transportation Tweets in 2022 (left) and Deviation of Share from Average of All Years (right)**

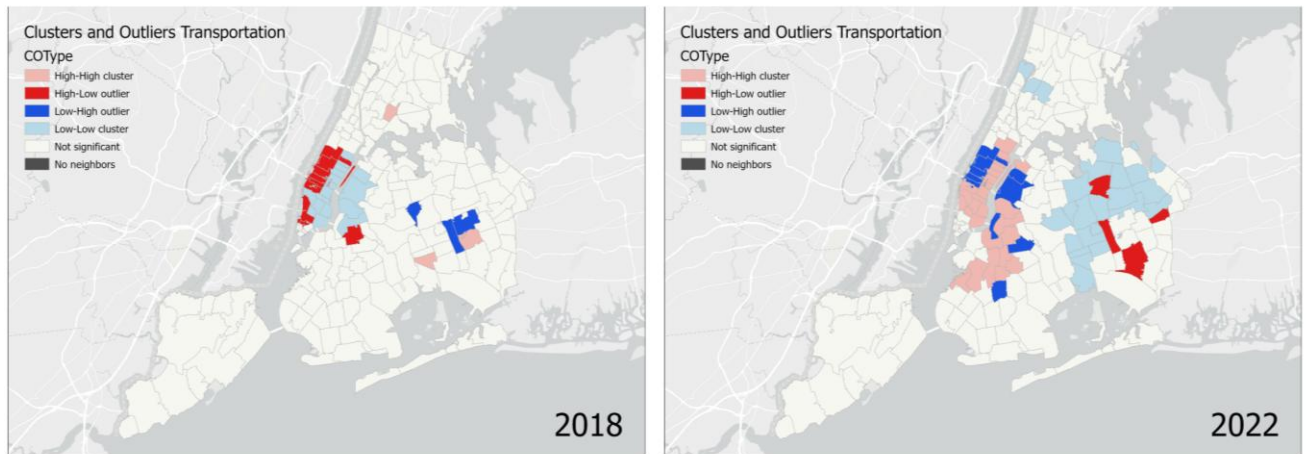
To statistically evaluate whether the observed deviations are spatially clustered or random, Global and Local Moran’s I were applied to the deviation results, modeling the spatial relationships with a fixed-distance band. While 2018 already shows significant clustering, no significant spatial autocorrelation was detected for 2019 to 2021. By 2022 however, the clustering of deviations becomes highly significant (Table 3).

	2018	2019	2020	2021	2022
<b>Moran's Index</b>	0.082443	0.000329	0.027424	0.026214	0.21543
<b>z-Score</b>	3.019604	0.174993	1.10013	1.030507	7.505692
<b>p-value</b>	0.002531	0.861085	0.271276	0.302772	0

**Table 3 Global Moran's I Results of Distribution of Deviation of Transportation Shares over All-Year Average**

The Local Indicators of Spatial Association (LISA) maps (Figure 11) confirm that in 2018, significant clustering mainly reflects areas of consistently below-average transportation-related Tweet activity with stretches in eastern Manhattan and northern Queens complemented by high-low outliers in northeastern Manhattan and low-high outliers like Jamaica Station in Queens.

On the other hand, the 2022 LISA results underline the observed shift, with coherent high-high clusters along subway and railway lines in Manhattan and Brooklyn, while extensive low-low clusters dominate Queens. Outliers such as Jamaica Station (high-low) and selected ZIPs in Manhattan and Queens (low-high) remain.

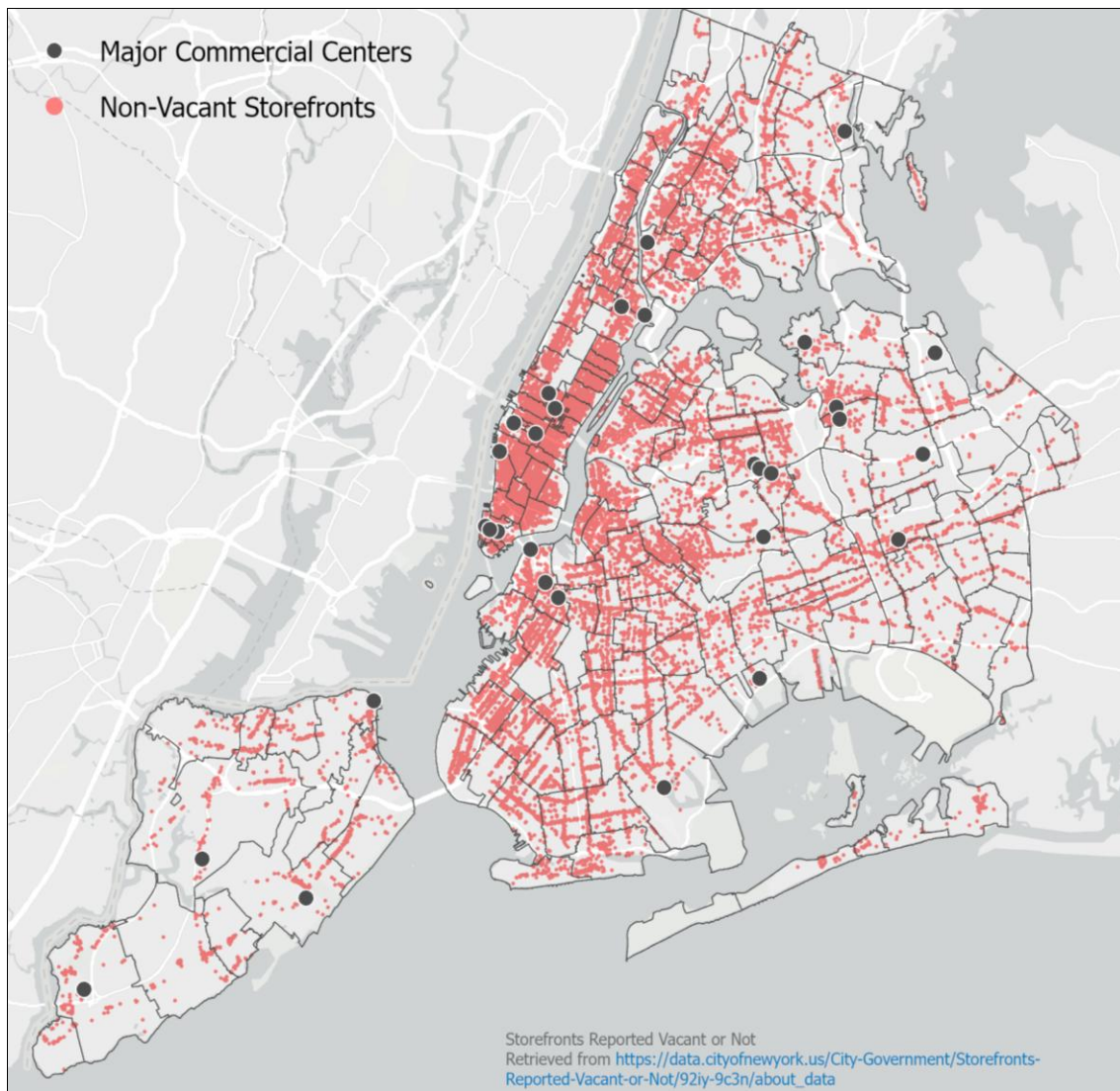


**Figure 11 LISA Maps of Transportation Deviation from All-Year Average in NYC**

#### 4.2.3 Retail Activity

Tweets related to Retail Activity appear in diverse forms, ranging from shared locations (“I’m at Union Square Greenmarket - @unsqgreenmarket in New York, NY”), marketing communication of special events (“Opening Night! Let’s take a swig! ? #brewery #openingnight #party #beer @ The Bronx Brewery”), business hours (“Toro NYC will be CLOSED this evening for a private event! We will reopen for dinner service”), direct product promotions (“3 Prong Tennis Chains Available Email Us [email address redacted] or call/text [phone number redacted]”) as well as human experiences with retail services or products (“Much needed haircut. (@ Asia Barber Shop in New York, NY)”; “I love a happy customer. Thank you! . . . naobeid #thankyou #interiordesign #interiordesigner”; “Brilliant market for halal food (@ Honest Chops Burgers in New York, NY”).

Figure 12 provides a brief overview of New York City’s retail infrastructure to support the interpretation of the subsequent shares and deviation results. Non-vacant storefronts are widely distributed across boroughs, typically tracing along shopping corridors, with the highest concentrations visible in Manhattan. In addition, major shopping centers are shown as reference points.



**Figure 12 Non-Vacant Storefronts and Major Commercial Centers in NYC (Department of Finance (DOF), 2025)**

Against this structural overview, the spatial distribution of Retail Activity related Tweet activity appears widespread across all boroughs but with varying spatial intensities over time (Appendix 6).

In 2018, shares of retail-related Tweets are broadly dispersed, showing a few isolated high share areas in Brooklyn and Queens while their respective deviations paint a mainly patchy picture with below-average exceptions in Brooklyn and Staten Island.

By 2019, shares remain broadly distributed but show slight increases in southern Brooklyn and northeastern Queens (Figure 13). The latter forms an above-average cluster despite containing comparatively few large shopping centers (like Bay Terrace Shopping Center or Fresh Meadows Shopping Center) or storefronts, which might suggest that retail-related Tweet activity is not solely tied to major malls and commercial corridors, but also to neighborhood scale retail or marketing operations. Parts of Staten Island also record higher shares and above-average deviations,

particularly around commercial areas such as Staten Island Mall in the northwest or Hylan Boulevard in the southeast.

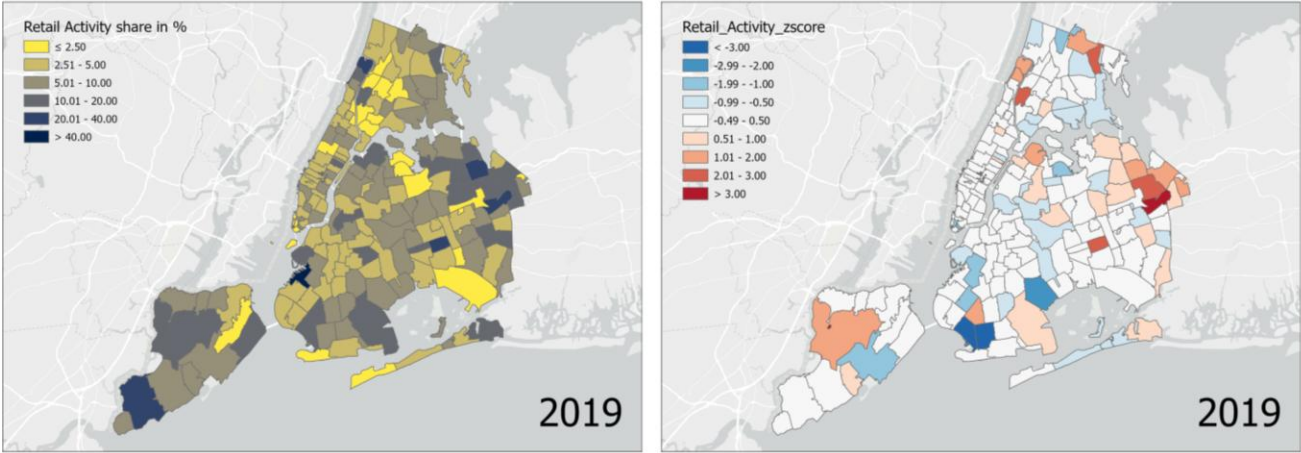
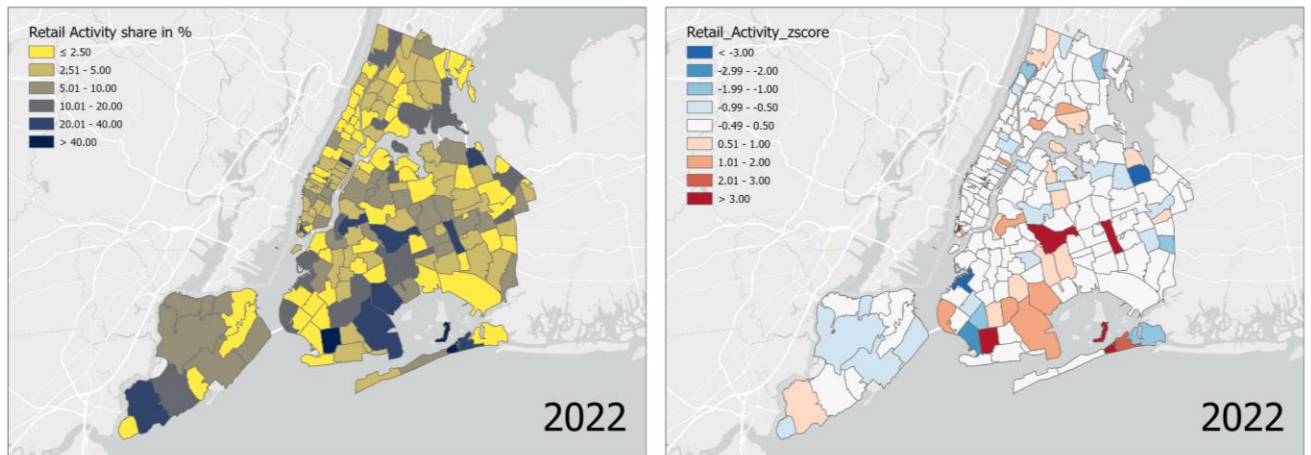


Figure 13 Share of Retail-Activity Tweets in 2019 (left) and Deviation of Share from Average of All Years (right)

In 2020 and 2021, retail-related Tweet activity continues to appear widespread across the city, but deviations show more ambivalent patterns. Some ZIP codes in southern Brooklyn and Staten Island record above-average values and high shares in proximity to larger commercial centers like King’s Plaza in Brooklyn or Staten Island Mall, yet these are interspersed with average or below-average results, so no stable cluster can be seen clearly. Overall, the two years mark a phase with local increases and only slight hints towards directional shift that, however, can not be directly linked to any COVID-19 zoning measures or other restrictions.

By 2022, the spatial structure becomes a bit more distinct and high share ZIP codes are especially grouped in southeastern Brooklyn close to the King’s Plaza and Gateway Center, which also stands out with above-average retail-related Tweet activity, while other boroughs like Manhattan, the Bronx or Queens show more mixed patterns in shares and deviations (Figure 14). Staten Island remains largely a borough with generally higher shares of retail-related Tweets, showing slight below-average values in 2022 while it is above-average in the southwestern in the area of the Bricktown Center at Charleston.



**Figure 14** Share of Retail-Activity Tweets in 2022 (left) and Deviation of Share from Average of All Years (right)

Even though the observed deviations visually suggested clusters in different areas and years in New York City, Global Moran’s detected significant spatial autocorrelation only for the year 2019 (Table 4), which might confirm the visible pattern in northeastern Queens, where deviation values temporarily stand out from the average baseline (Figure 13). However, the LISA analysis did not reveal any significant local clusters and therefore, the pattern remains limited to broader global tendencies than on a neighborhood level.

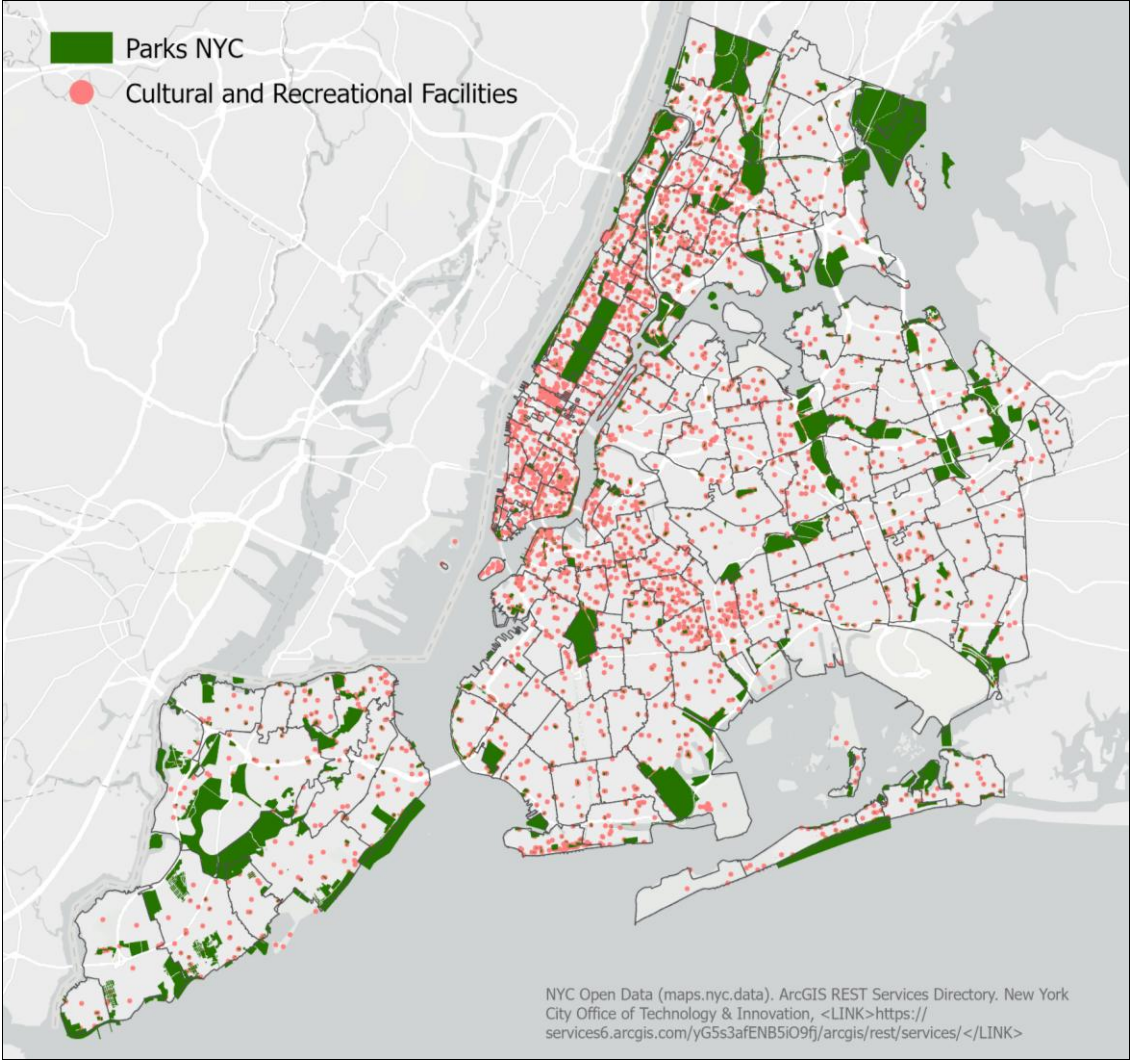
	2018	2019	2020	2021	2022
<b>Moran's Index</b>	0.037496	0.110334	-0.02103	0.027659	0.03555
<b>z-Score</b>	1.425845	3.940511	-0.548374	1.136133	1.393332
<b>p-value</b>	0.153913	0.000081	0.583435	0.255901	0.163519

**Table 4** Global Moran's I Results of Distribution of Deviation of Retail-Activity Shares over All-Year Average

#### 4.2.4 Cultural/Social Activity

Tweets categorized as Cultural/Social Activity encompass a wide range of practices, from shared cultural locations or photos (“I’m at New York Theatre Workshop - @nytw79 in New York”; “Just posted a photo @ Times Square, New York City”), to public events (“NYC is ready for #NYE2023 @ Times Square, New York City”), various kinds of leisure (“Me enjoying a hearty bowl of onion soup; loaded with onions and melty cheese atop toasted baguette.”; “Zoo day with #NaiaNickz @ The Bronx Zoo”), community or creative activities (“Do You Love to Write? Join the LiveJournal Community and Write on Your Own Blog!”; “Giving back. Spent the morning as a volunteer in NYC packing meals for home bound people. Gods Love We Deliver prepares and delivers 7,500 meals a day! A pleasure to help out with my Douglas Elliman colleagues.”) to everyday life situations and thoughts (“Happy Sunday everyone! #sundayfunday #myboy #?? @ Briarwood”; “Christmas Day. Home. (@ [address redacted] in New york, NY”).

Even though cultural-/social-related Tweet content is broadly themed, Figure 15 illustrates the distribution of parks and designated cultural and recreational facilities in New York City. The facilities are dispersed across all boroughs but show strong concentrations in Manhattan and dense clusters along major corridors in Brooklyn, Queens, and the Bronx. Staten Island, in contrast, is characterized by a larger share of parkland and more dispersed cultural facilities.



parts of Manhattan. The central Bronx also shows above-average activity, however, this effect may be partly driven by a specific, highly active account sharing public lots and parks, which might therefore not reflect broad communal activity.

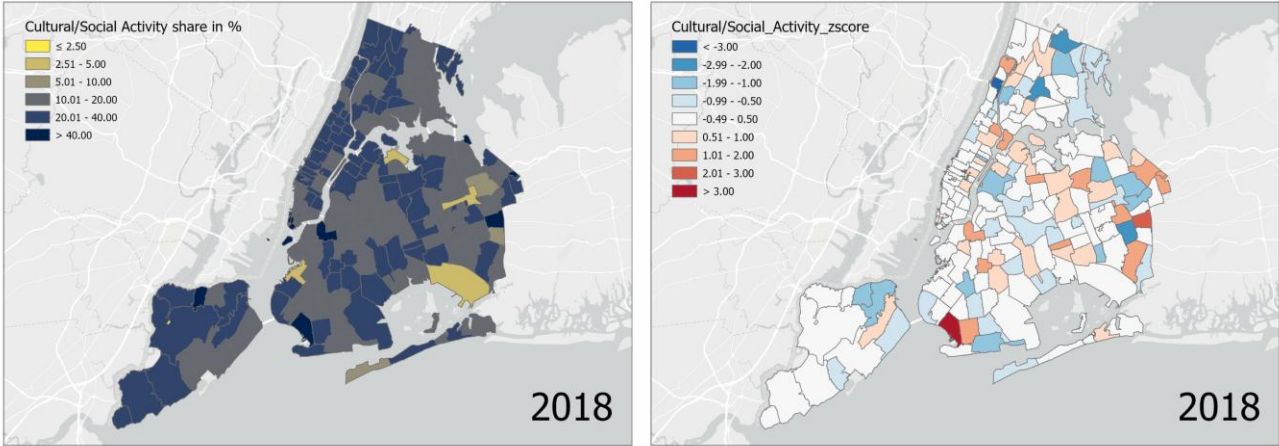


Figure 16 Share of Cultural/Social Activity Tweets in 2018 (left) and Deviation of Share from Average of All Years (right)

In 2020, the spatial distribution changes more distinctly (Figure 17). While shares remain widespread, an above-average corridor emerges between Brooklyn and Queens, contrasting with surrounding areas where the deviation values drop below the long-term average.

This pattern partly borders COVID-19 restriction zones (Appendix 8) introduced in autumn 2020 (Cuomo, October 21, 2020, Harris, 2022), suggesting that pandemic-related regulations may have reshaped cultural and social activity in these neighborhoods. However, it also partially aligns with the higher concentration of cultural and recreational facilities visible in Figure 15.

Additional above-average Tweet activity can be observed in parts of Staten Island containing larger park areas and along the eastern shoreline of the Bronx close to larger park areas.

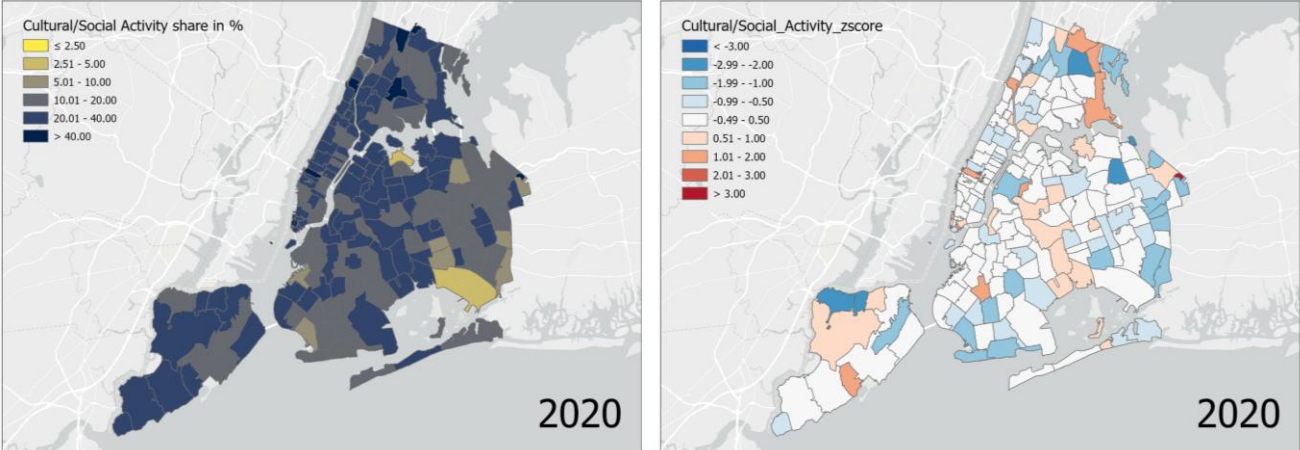


Figure 17 Share of Cultural/Social Activity Tweets in 2020 (left) and Deviation of Share from Average of All Years (right)

In 2021, cultural and social Tweet activity remains widespread, with above-average values particularly visible in Staten Island, the Jamaica area, and parts of northeastern and central Queens, where recreational facilities usually are less concentrated and larger park areas available. Shares also seem elevated along the Rockaway Peninsula.

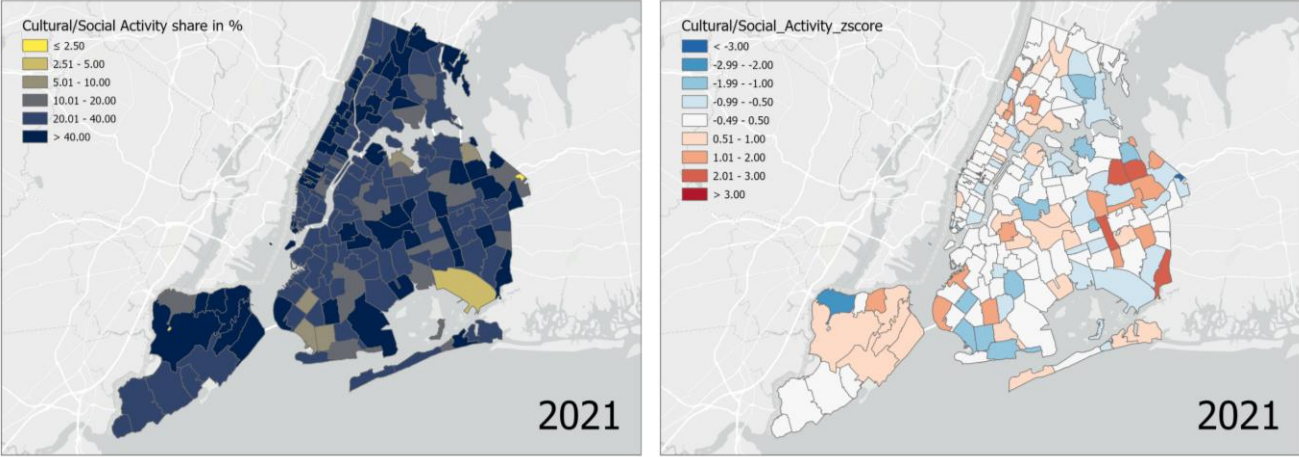


Figure 18 Share of Cultural/Social Activity Tweets in 2021 (left) and Deviation of Share from Average of All Years (right)

By 2022, overall shares stay high across the city, but deviations become more mixed and no coherent spatial pattern emerges except for above-average area visible in central Queens. This pattern may reflect the gradual return to pre-pandemic diversity.

The results of Global Moran’s indicate significant clustering for the observed deviations in the years 2019, 2020 and 2022, while 2018 and 2021 do not show statistically significant autocorrelation (Table 5), even though few visual clusters were assessed in 2021 (Figure 18).

	2018	2019	2020	2021	2022
<b>Moran's Index</b>	-0.0039	0.053107	0.067381	0.005039	0.094271
<b>z-Score</b>	0.029215	1.967775	2.409147	0.328138	3.30508
<b>p-Value</b>	0.976693	0.049094	0.01599	0.742807	0.000949

Table 5 Global Moran's I Results of Distribution of Deviation of Cultural/Social-Activity Shares over All-Year Average

For the year 2020, LISA provides further results (Figure 19) that partly diverge from the deviation patterns seen in Figure 17. In Manhattan and parts of northwestern Queens, many ZIP codes are identified as low-high outliers or high-high clusters, even though their deviation values are mostly average, which reflects more their contrast towards adjacent areas rather than strong global deviations.

Staten Island, by contrast, shows a clearer alignment, with above-average deviations coinciding with high-low outliers, underlining localized intensifications in the northwest.

Around Jamaica in central Queens, a mix of high-low outliers and adjacent low-low clusters points to localized contrasts, while Coney Island in south Brooklyn forms a low-low cluster consistent with its below-average deviation.

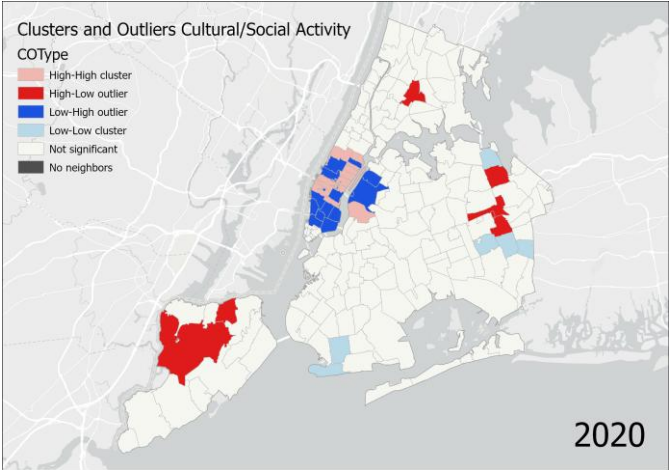


Figure 19 LISA Maps of Cultural/Social Activity Deviation from All-Year Average in NYC

## 5. Discussion

### 5.1 Discussion of Methodology

The development and evaluation of this workflow directly addresses the methodological research question of this study: **(RQ 1)** *How can geotagged Twitter posts be systematically filtered and classified into the urban life categories of Transportation, Retail Activity, Cultural/Social Activity, Healthcare and Work/Remote Work?*

The combination of keyword filtering and BERTopic-based clustering was designed to operationalize this question by creating a reproducible framework for extracting, structuring, and analyzing urban functional categories across multiple time periods. The following discussion therefore focuses on the representativeness, scalability and limitations of this approach.

The methodological approach developed in this thesis achieved precision scores above 90%, indicating that the combination of keyword filtering and BERTopic-based classification provides a relatively robust framework for capturing urban functional categories in geotagged Twitter data. However, these results should be interpreted with caution as precision partly reflects the manual mapping of BERTopic topics to those higher level categories and the trade-offs made during data pre-processing.

A central constraint arises from the sparsity of geolocated Tweets, which represent only a fraction of the overall available Twitter data and therefore demands for a cautious interpretation of visible patterns and results as the citywide communication is not reflected in its entirety. This, as well as the user demographics, tend to skew the results towards a digitally active minority and are widely noted aspects in current research (De Sabbata et al., 2023, Arifi et al., 2025, Kovacs-Györi et al., 2018, Yang and Liu, 2022, G. Almatar et al., 2020). Expanding the data by retrieving locations from the Tweet texts like park names in the work of Huang et al. (2022) or combining geotagged with self-reported locations (G. Almatar et al., 2020) might enhance spatial coverage.

In addition to spatial representativeness, linguistic coverage represents another limitation. The workflow could, in principle, be extended to multiple languages as BERTopic supports multilingual input (Egger and Yu, 2022). However, due to the hybrid nature of the workflow, additional adjustments would be required for translating text and keywords into the target language or curating equivalent keyword lists for each language directly. While large language models like BERTopic already are computationally intensive when used for large datasets (Arifi et al., 2025), authors like Jiang et al. (2023) underline the extension of BERTopic to different languages as a computational challenge.

A related aspect of representativeness involves the detection of non-individual accounts, including both: automated bots and organizational users. Even if the chosen heuristic approach was simple and effective towards the removal of obvious organizational accounts, it may have overlooked more subtle automated or semi-automated user accounts. Similar concerns are raised by Ferrara et al. (2016), who address that social bots can distort online social media activity by amplifying noise and misinformation and make it more difficult to distinguish between bot and human-like behavior. Therefore, residual automated users may influence spatial and thematic patterns observed in the present dataset. Future research could explore more adaptive detection approaches (Ferrara et al., 2016) or treat organizational accounts separately instead of removing them entirely and acknowledge their role as institutional actors in urban communication (McCorrison et al., 2015).

Beyond spatial, linguistic and user representativeness, further methodological challenges emerged. The keyword extraction process required continuous tuning of the parameters to maximize the topic relevance of the resulting keywords while also minimizing noisy terms that would require more manual correction afterwards. This was also addressed by De Sabbata et al. (2023) and G. Almatar et al. (2020), who emphasize that topic models on social media data often require iterative manual refinement to preserve conceptual coherence. This interpretative component also extends to the manual mapping of BERTopic-generated topics to the predefined categories such as Retail Activity, Healthcare, Transportation, Work/Remote Work and Cultural/Social Activity. Even though the manual mapping allowed for a more precise classification of the results, automation might enhance the reproducibility of the process.

Also, to prevent overlapping assignments across categories, BERTopic was only applied after the keyword filtering step, meaning that Tweets entering the model already contained category-relevant keywords. The design choice used BERTopic as a final assignment step towards the urban categories, grouping already relevant Tweets into coherent subclusters. However, a substantial share of these filtered Tweets were assigned to the “-1” outlier topic as they could not be meaningfully clustered despite their categorical relevance. This side effect was also reported by authors like De Sabbata et al. (2023), who state that a considerable portion of the Twitter data remained unclustered in their BERTopic application due to thematic ambiguity and by Steiger et al. (2016) who observed that sparse or heterogeneous social media text can hinder the formation of coherent semantic clusters. While this mechanism is valuable for filtering out incoherent text or weak signals, it also results in the loss of Tweets that were thematically relevant but too sparse or too heterogeneous to form stable clusters. Consequently, resulting category distributions may underrepresent less frequent but possibly still meaningful aspects of urban Tweet activity.

The final spatial aggregation to ZIP code tabulation areas was chosen, as these units contained enough Tweets for spatial analysis, whereas smaller units caused greater inconsistencies due to sparse coverage. This is consistent with observations that Tweet based spatial densities vary significantly by choice of spatial units (Jiang et al., 2016) and aligns with findings by Li et al. (2021), who examined place connectivity across multiple spatial scales using geotagged social media data.

To ensure comparability across areas with uneven posting volumes, Tweet frequencies were normalized by the total number of Tweets per ZIP code and analyzed as deviations from each area's long-term mean. This approach parallels the z-score standardization and baseline deviation analysis employed by Kontokosta et al. (2024), who examined neighborhood-level changes in social media attention by quantifying fluctuations from each area's long-term popularity average.

Finally, the results are validated through spatial correspondence with ground truth facilities and infrastructures rather than official statistics. For example, transport-related patterns were compared to the locations of public transport lines and hubs or retail-related patterns to commercial center locations. This approach ensured that observed patterns remained anchored in the urban landscape, comparable to observations by Steiger et al. (2016) who detected links between activities related to sport events and major transport hubs or Frias-Martinez et al. (2012) and Lansley and Longley (2016) who validated Tweet-derived activity clusters against official land use data and landmark locations in New York City and London.

## **5.2 Discussion of Results**

The following section addresses the research questions **(RQ 2)** *How did categorical Tweet activity in New York City change between pre-pandemic, pandemic and post-pandemic periods?* And **(RQ 3)** *How do the spatial distributions of categorical Tweet activity change across pre-pandemic, pandemic, and post-pandemic periods, and where do localized intensifications or declines become visible?*

Building on the previously validated methodological framework **(RQ 1)**, this chapter interprets the spatiotemporal patterns of the classified Tweet data across thematic domains. It examines how contractions, recoveries and redistributions of categorical Tweet intensity correspond to observable urban structures and possible pandemic related constraints and relates these findings to existing literature.

The analysis of geolocated Twitter data in New York City between 2018 and 2022 reveals distinct temporal and spatial dynamics across various thematic categories, providing insights into the evolving urban Tweet activity in response to external events such as the COVID-19 pandemic. While some themes like Transportation and Cultural/Social Activity demonstrate clear patterns of contraction and recovery, others, particularly Retail Activity, present more nuanced and localized

fluctuations. Healthcare and Work/Remote Work, conversely, exhibit limited and patchy signals, suggesting that their geotagged Tweet volumes might be too limited for robust spatial and temporal patterns. This imbalance is also underlined in Liao et al. (2022), who acknowledge clear indications of leisure activities being overly represented in social media data while also being a viable source to quantify mobility patterns (Liao et al., 2022, Jurdak et al., 2015).

Pre-pandemic transportation-related Tweet intensity was geographically more scattered, with even Manhattan showing below-average activity, potentially due to dense, non-digital communication or data saturation. The pandemic year shows reduced activity, while transportation hubs like Jamaica Station illustrate higher Tweet shares. Similar temporal contractions in transport-related social media activity during early pandemic phases were observed by Rajput et al. (2022), Wang et al. (2021) and Jiang et al. (2021), who reported clear declines, especially in public transport discussion in major cities as mobility restrictions took effect.

By 2022, a highly significant clustering emerged, with high-high clusters along subway and railway lines in Manhattan and Brooklyn, possibly demonstrating a re-establishment of public transit as topic of interest in digital conversation. This shift from scattered activity to concentrated clusters might highlight a post pandemic return to physically anchored transit discussions. It also aligns with findings from authors like Zhong et al. (2023) and Forouhar et al. (2025), who record recovery patterns in mobility-related social media data in proximity to transit stations during the post pandemic years in cities like London and Toronto. Regarding New York City, Qiang and McKenzie (2024) underline the return of public transport activity by pandemic phase and emphasize that this return is not uniform and greatly affected by the role of anti-pandemic policies and socio-economic characteristics of the region. Further addressing these diverging circumstances could be part of future research regarding transportation-related social media activity.

The Cultural/Social Activity category displayed a widespread spatial footprint before the pandemic, underscoring its central role in urban life as noted by Liao et al. (2022) and Jurdak et al. (2015). In 2020, during pandemic restrictions, the activity pattern exhibited a marked spatial contrast, with a pronounced above-average corridor between Brooklyn and Queens, which partly borders areas that were subject to COVID-19 restriction zones and coincide with dense concentrations of cultural venues and facilities. While existing studies have addressed the limited effectiveness of these zones in containing the virus (Harris, 2022), cultural or social activity patterns in these areas have not yet been examined. Nevertheless, the spatial correspondence may suggest that localized engagement with cultural and social themes persisted in this corridor despite the broader disruption of urban activities.

At the same time, parts of Staten Island and the eastern shoreline of the Bronx also notably displayed above-average cultural/social Tweet intensity during 2020, particularly areas containing larger parklands, indicating a potential shift towards outdoor activities during restrictions. This is consistent with the work of Zhao et al. (2023), who found that nature areas in New York City experienced stable or even increased visitation during the pandemic and served as perceived safe havens for residents, particularly in outer-borough areas. Their results suggest that shifts in recreational behavior toward peripheral parks reflected a broader adaptation to various implemented restrictions to combat the spread of the disease which might explain the results seen in this study.

At this point, statistically significant clusters were only recorded for the regions in Staten Island, that likely reflect the borough's large parklands and lower density, which provided accessible outdoor venues for social and cultural activities during pandemic restrictions (Zhao et al., 2023) while the observed corridor in Brooklyn and Queens and the eastern shoreline of the Bronx did not display statistically significant clustering. Following the, possibly pandemic-induced, clustering of deviations, while overall shares remained high, the post-pandemic years saw a gradual return to mixed patterns. These findings align with previous observations that cultural and creative social media activity is often less spatially clustered and more widely distributed across urban landscape (Reuschke et al., 2021, Niu and Silva, 2023), reflecting the more diverse and decentralized nature of cultural engagement in larger cities.

In contrast, Retail Activity Tweets exhibited more fragmented and less spatially coherent patterns. While 2019 showed global significant clustering of retail-related Tweet activity, the LISA analysis did not identify coherent local clusters in the visually assessed above-average region in northeastern Queens. Similarly, localized increases in southern Brooklyn 2022 near commercial centers such as King's Plaza are not supported by significant clustering and remain a visual pattern. This overall heterogeneity may reflect the diverse nature of retail-related social media interactions, which range from consumer reviews to business promotions and vary strongly in tone and purpose. Li et al. (2023) demonstrated that customers engage in different kinds of social media interactions during crises, each representing different attributes, resulting in uneven online activity across businesses. These different ways of engagement likely contributed to the spatial fragmentation of retail-related Tweet activity observed in this study. Future research could therefore benefit from distinguishing between consumer- and business-generated retail Tweets.

Taken together, the results suggest that the interpretative value of geotagged Tweets depends strongly on the thematic context and the density of the data. Only in thematic domains like Transportation or Cultural/Social Activity, where sustained engagement occurs among an already digitally active user base, can spatial and temporal tendencies be recognized, while other themes

remain too sparse for meaningful interpretation. Analyzing deviations from long-term baseline category shares proved useful for detecting those tendencies in form of spatially clustered disruptions associated with pandemic-related changes, particularly in domains such as transportation and cultural/social activity. The observed spatial alignments therefore illustrate not comprehensive urban behavior, but the digital activity of selected regions and how they react to structural changes in urban life due to the COVID-19 pandemic.

## **6. Conclusion**

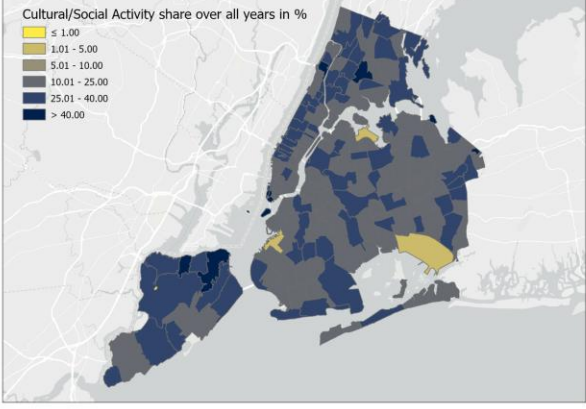
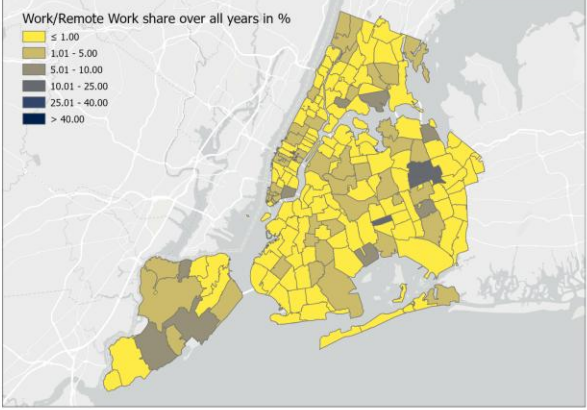
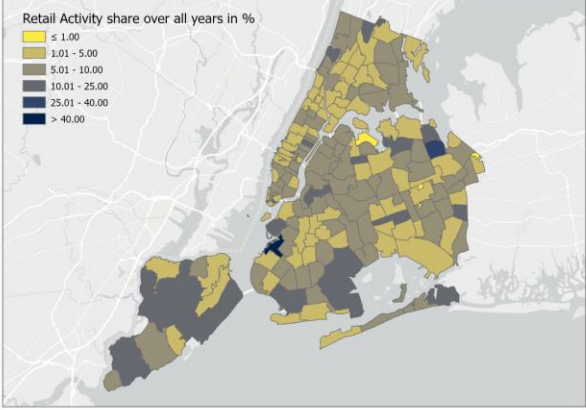
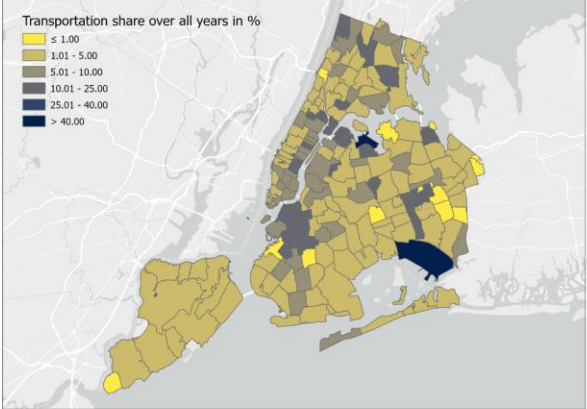
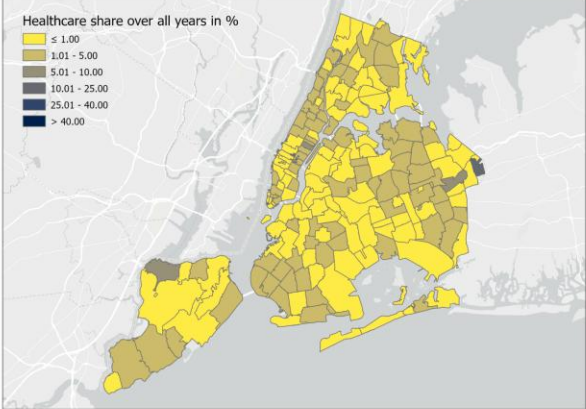
This thesis explored how geotagged Twitter data can be used to analyze thematic Tweet activity related to urban functional categories and how such activity in New York City changed between pre-pandemic, pandemic, and post-pandemic periods. Addressing the first research question, the developed workflow illustrates how keyword filtering combined with BERTopic-based classification can be applied to filter, structure and interpret category-related content from large-scale Twitter data in a reproducible manner.

In relation to the second and third research questions, the temporal and spatial analysis revealed that deviations from long-term baseline shares can highlight spatially clustered disruptions and recovery phases in category-related Tweet activity, particularly within transportation and cultural/social activity categories. These results align with existing literature emphasizing the sensitivity of social media signals to events in the real world.

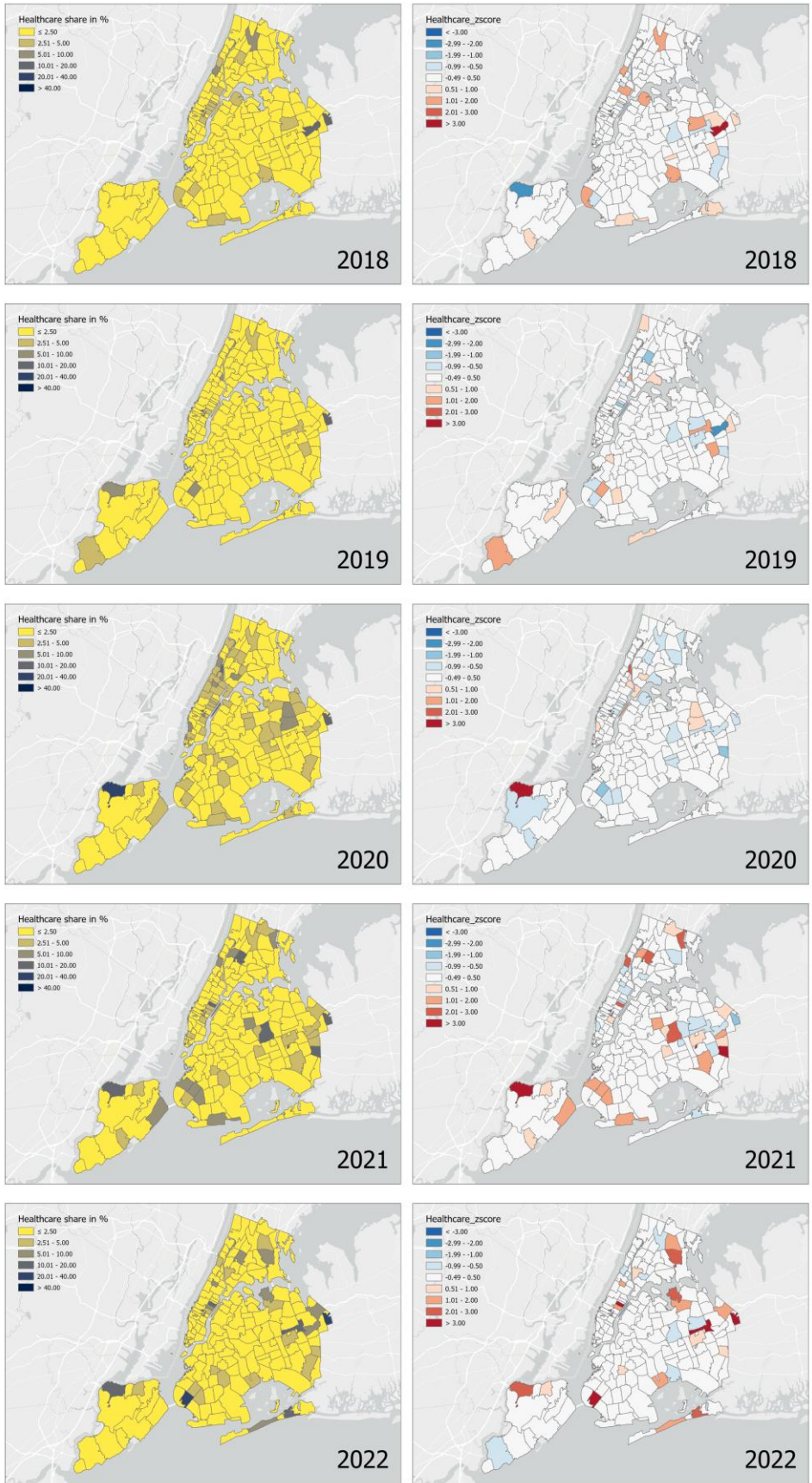
While data sparsity, outlier exclusion and manual interpretation constrain the completeness and reproducibility of the findings, the workflow nonetheless demonstrates potential for identifying thematic changes in categories that were sufficiently represented. For these more prominent themes, deviations corresponded plausibly with features of the city's physical infrastructure, whereas other categories remained too fragmented for robust interpretation.

Rather than representing comprehensive urban behavior, geotagged Twitter data reflects the perspectives of a digitally active minority, yet it responds sensitively to major disruptions and changes in urban life. The developed approach therefore offers a complementary and exploratory means of tracing selected aspects of urban social media activity over time, particularly during periods of crisis or transition. Future work could expand this approach across multiple cities or integrate multilingual and/or inferred location data to enhance coverage and generalizability.

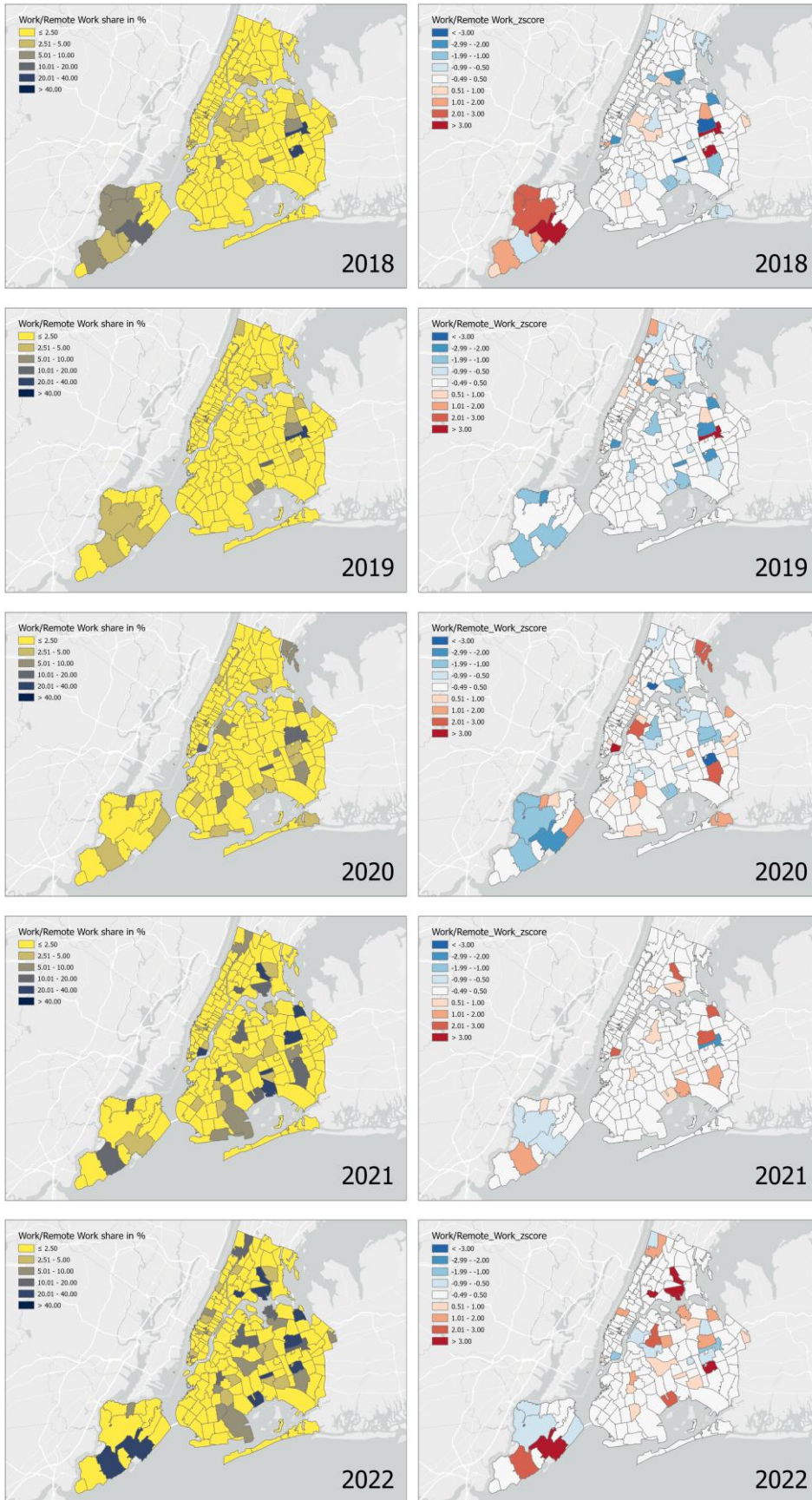
# Appendices



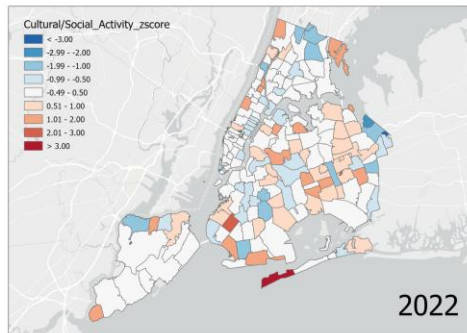
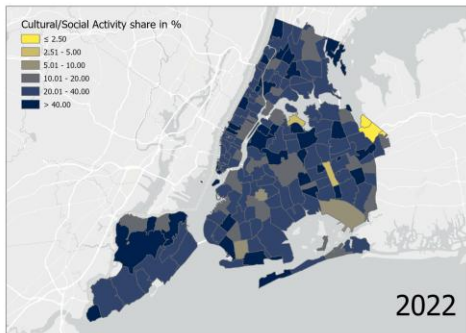
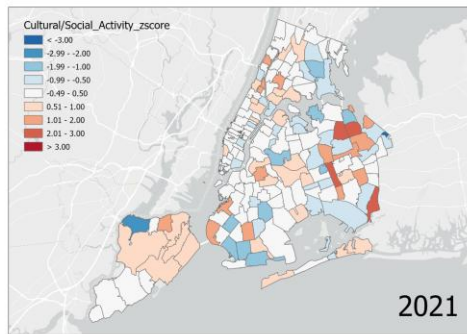
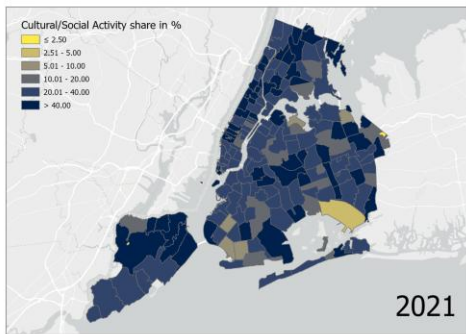
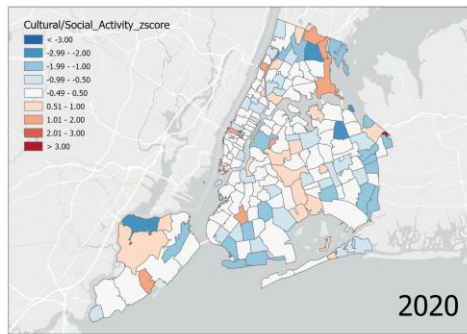
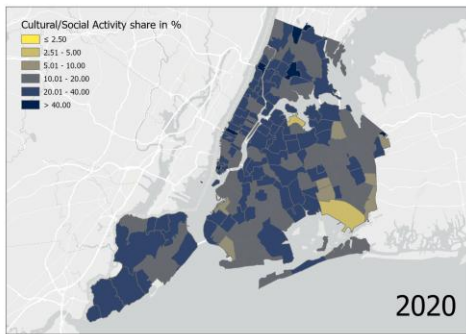
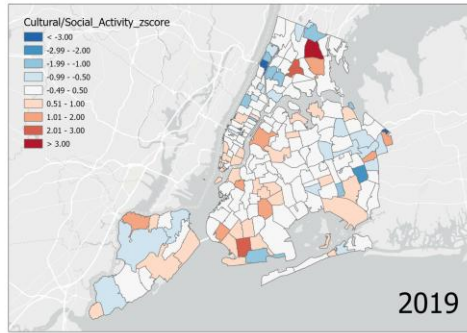
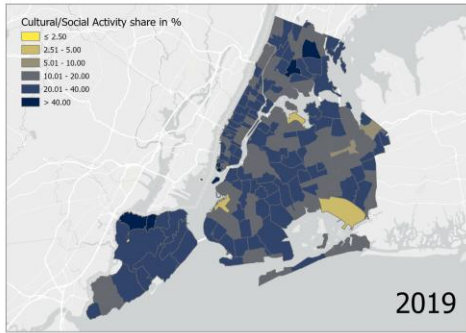
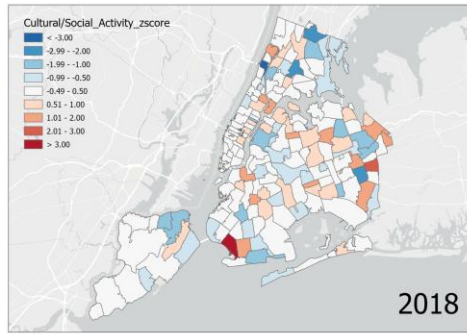
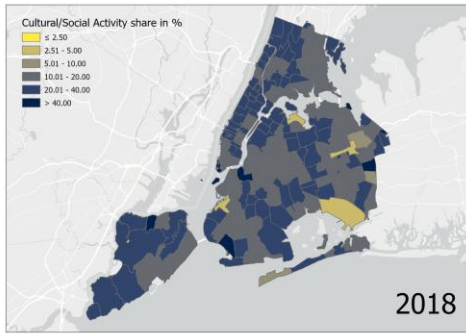
Appendix 1 All-Year Average Shares per Category in NYC



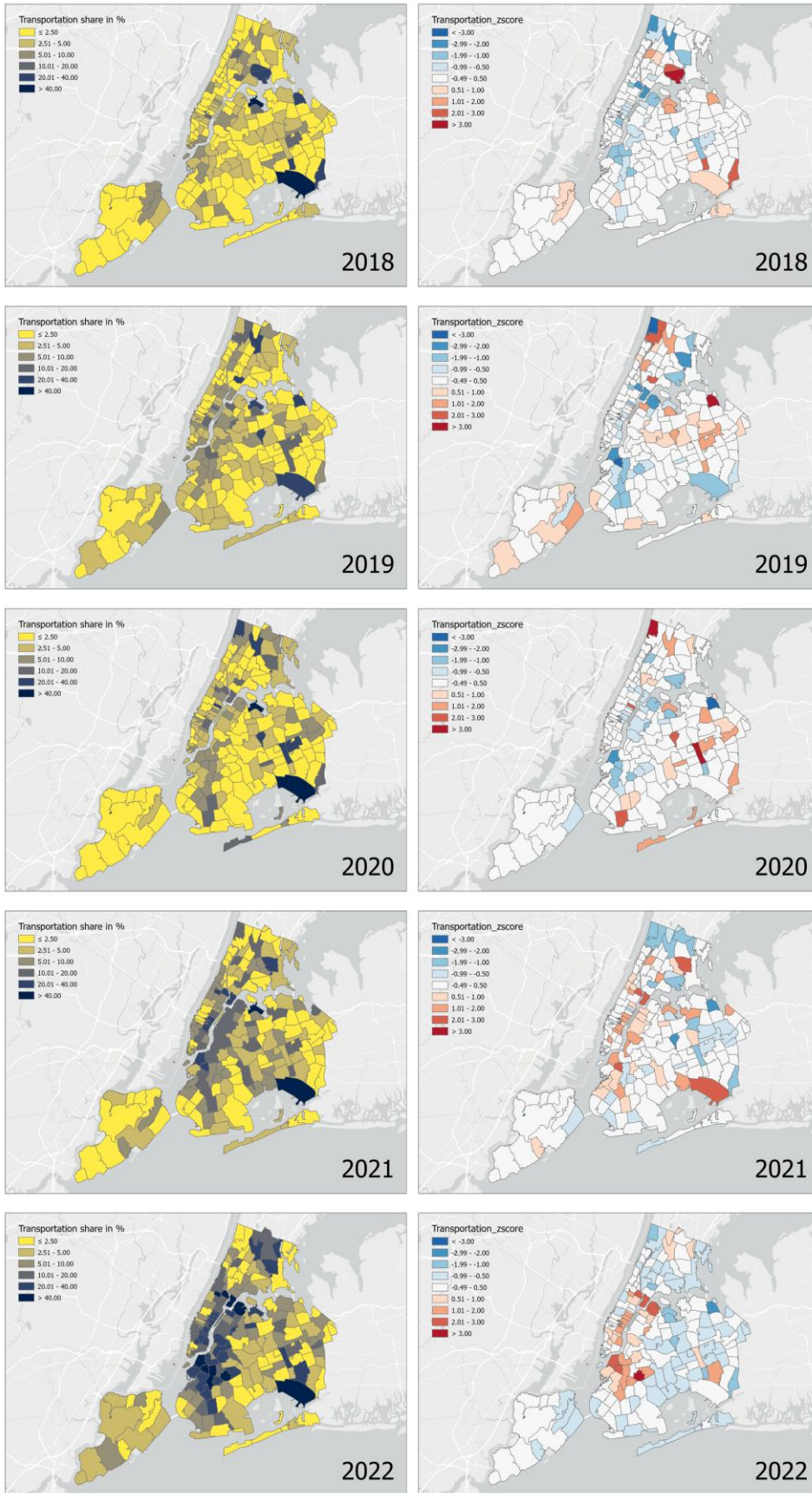
Appendix 2 Shares of Healthcare Category (left) and Deviation of Shares from All-Year Average in NYC (2018-2022)



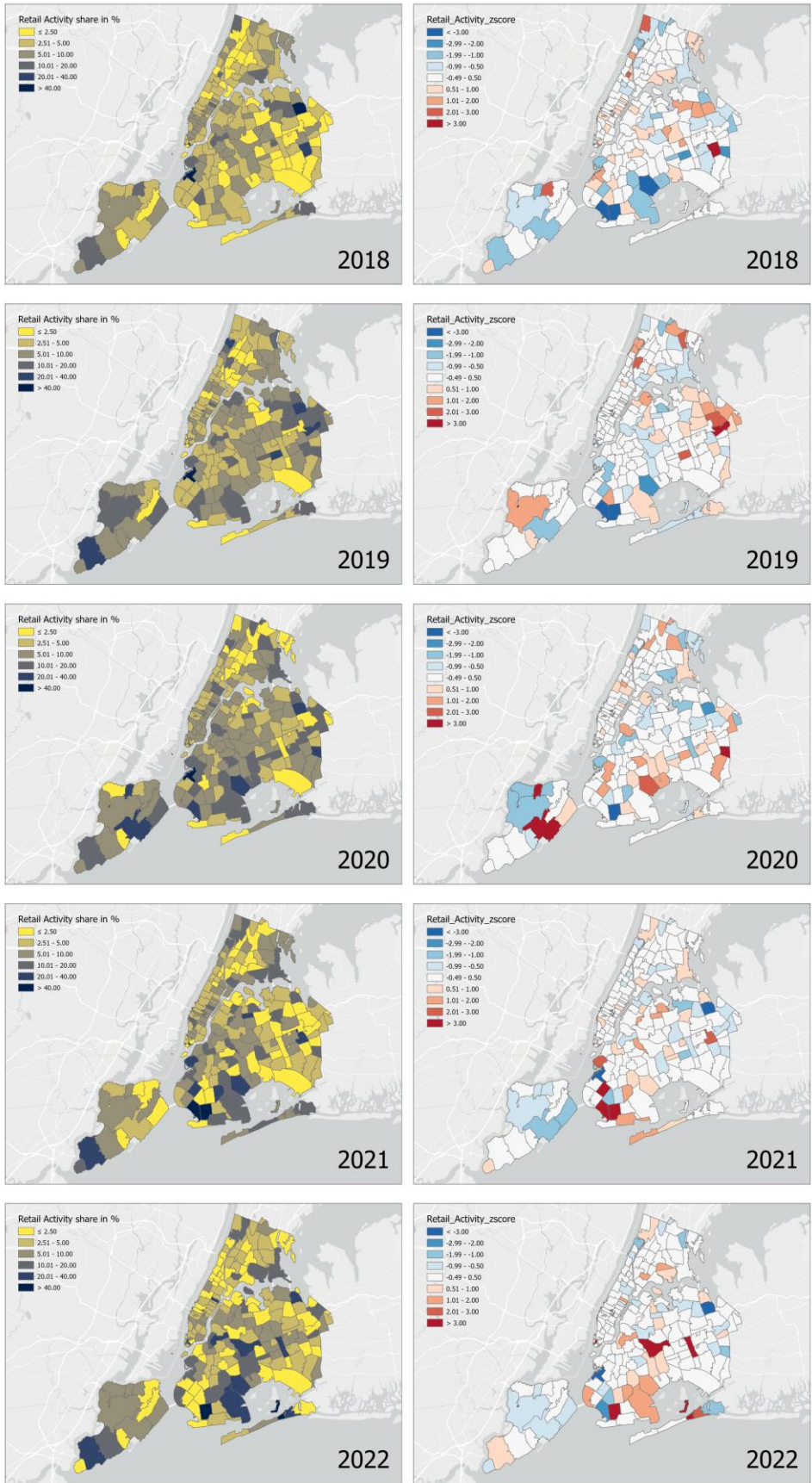
**Appendix 3 Shares of Work/Remote Work Category (left) and Deviation of Shares from All-Year Average in NYC (2018-2022)**



Appendix 4 Shares of Cult./Social-Activity Category (left) and Deviation of Shares from All-Year Average in NYC (2018-2022)



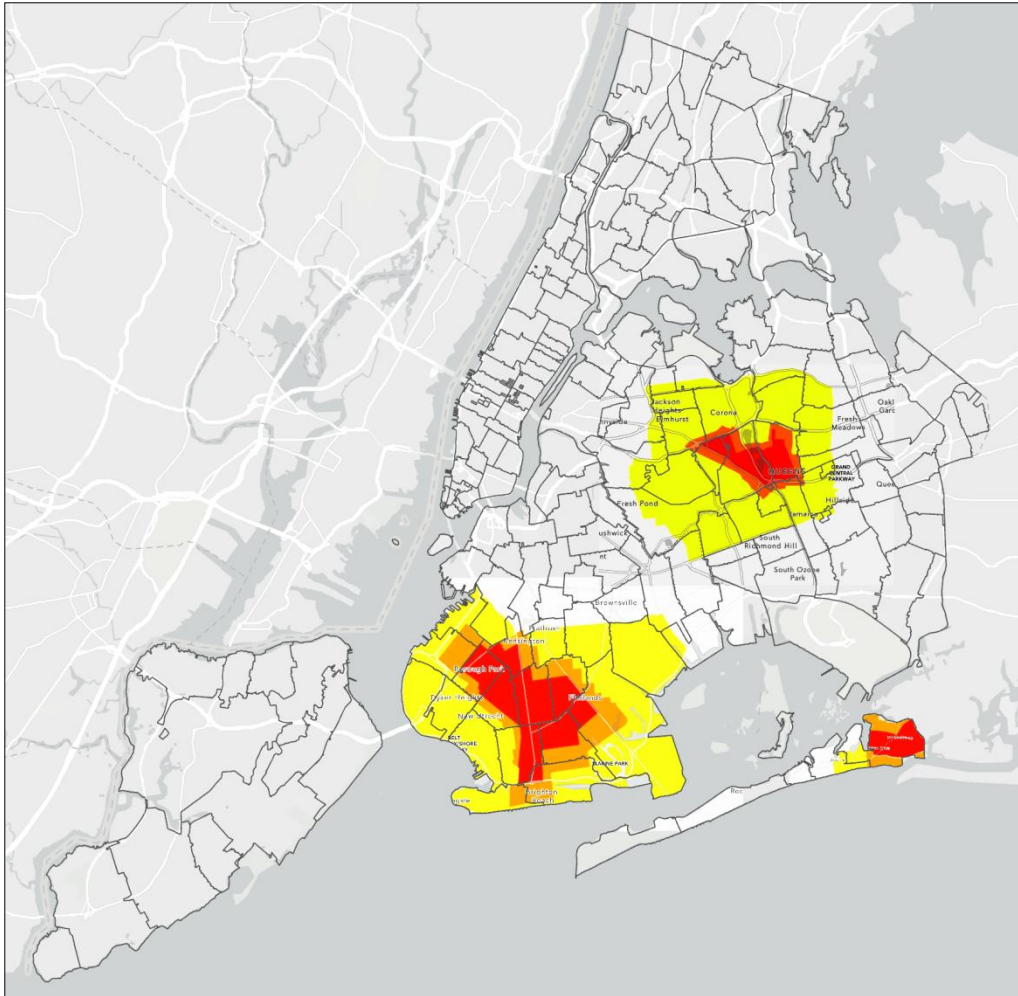
Appendix 5 Shares of Transportation Category (left) and Deviation of Shares from All-Year Average in NYC (2018-2022)



Appendix 6 Shares of Retail-Activity Category (left) and Deviation of Shares from All-Year Average in NYC (2018-2022)

		2018	2019	2020	2021	2022
<b>Healthcare</b>	<b>Moran's Index</b>	0.045616	0.012571	0.08286	0.017053	0.025729
	<b>z-Score</b>	1.866941	0.699192	3.137028	0.741009	1.058774
	<b>p-value</b>	0.06191	0.484432	0.001707	0.458688	0.289703
<b>Transportation</b>	<b>Moran's Index</b>	0.082443	0.000329	0.027424	0.026214	0.21543
	<b>z-Score</b>	3.019604	0.174993	1.10013	1.030507	7.505692
	<b>p-value</b>	0.002531	0.861085	0.271276	0.302772	0
<b>Retail Activity</b>	<b>Moran's Index</b>	0.037496	0.110334	-0.02103	0.027659	0.03555
	<b>z-Score</b>	1.425845	3.940511	-0.548374	1.136133	1.393332
	<b>p-value</b>	0.153913	0.000081	0.583435	0.255901	0.163519
<b>Work/Remote Work</b>	<b>Moran's Index</b>	0.052758	-0.016481	-0.016693	-0.008383	-0.029282
	<b>z-Score</b>	2.030953	-0.425373	-0.421837	-0.147424	-0.876876
	<b>p-value</b>	0.04226	0.670565	0.673144	0.882798	0.380554
<b>Cultural/Social Activity</b>	<b>Moran's Index</b>	-0.0039	0.053107	0.067381	0.005039	0.094271
	<b>z-Score</b>	0.029215	1.967775	2.409147	0.328138	3.30508
	<b>p-value</b>	0.976693	0.049094	0.01599	0.742807	0.000949

Appendix 7 Global Moran's I Results of Share Deviation from All-Year Average (All Categories)



**Appendix 8 "Cluster Zone" Focus Areas Designated for NYC (NYC Mayor's Office [@NYCMayorsOffice], 2020, Cuomo, October 21, 2020)**

**Appendix 9 Final Curated List of Cultural/Social Activity Wikipedia Articles**

"Entertainment", "Culture", "Society", "Drama", "Performance", "Court (royal)", "Banquet", "Party", "Amusement", "Fun", "Ceremony", "Religious festival", "Satire", "Recreation", "Leisure", "Play (theatre)", "Opera", "Television show", "Game", "Theatre", "Concert", "Magic (illusion)", "Video game", "Festival", "Music festival", "Film festival", "Competitive dance", "Spectator sport", "Cooking", "Remix", "Camping", "Circus", "Music hall", "Theater (structure)", "Auditorium", "Spectacle", "Racing", "Tournament", "Musical instrument", "Jazz", "Folk music", "Choir", "Musical ensemble", "Human voice", "A cappella", "Operetta", "Home cinema", "History of film", "Parade", "Fair", "Fandom", "Social relation", "Friendship", "Social phenomenon", "Social conflict", "Social roles", "Social behavior", "Social norm", "The arts", "Socialization", "Cultural norm", "Culture change", "UNESCO", "Anthropology", "Cultural universals", "Art", "Music", "Dance", "Ritual", "Religion", "Mythology", "Philosophy", "Literature", "Writing", "Oral literature", "Science", "Social class", "Popular culture", "Folk culture", "Cultural capital", "Media culture", "Tradition", "Cultural anthropology", "Cultural relativism", "Social interaction", "Cultural change", "Social innovation", "Social revolution", "Cultural invention", "Social structure", "European classical music", "Social group", "Western culture", "Archaeology", "Sociology", "Non-material culture", "Social stratification", "Social network", "Social psychology", "Social theory", "Film theory", "Museum studies", "Art history", "Film", "Photography", "Fashion", "Hairstyle", "Cultural artifact", "Culture shock", "LGBT culture", "Cultural heritage", "Cultural diversity"

## Appendix 10 Final Curated List of Work/Remote Work Wikipedia Articles

"Work\_(human\_activity)", "Job", "Remote\_work", "Office", "Coworking", "Job characteristic theory", "Job satisfaction", "Employee engagement", "Work-family conflict", "Turnover intention", "Work motivation", "Labour inspectorate", "Payment", "Unpaid work", "Manufacturing", "Corporation", "Profession", "Job titles", "Career", "Division of labor", "Critique of work", "Bullshit Jobs", "The Abolition of Work", "Retirement", "Universal basic income", "Employment", "Night shift", "Slavery", "Working class", "Hours of work", "Child labour", "Office work", "Seniority", "Apprentice", "Journeyman", "Master craftsman", "Skilled trade", "Managerial", "Manual labor", "Wage labor", "Piece work", "Overwork", "White-collar worker", "Sedentary", "Handicraft", "Computer (job description)", "Labor union", "Labor market", "Corporate", "Resource allocation", "Work ethic", "Protestant work ethic", "Slave labor", "Human trafficking", "Unemployment", "Productivity", "Job matching", "Unemployment insurance", "Job guarantee", "Cheap labour"

## Appendix 11 Final Curated List of Retail Activity Wikipedia Articles

"Retail", "Goods", "Service (economics)", "Consumer", "Wholesaling", "Institutional customers", "Manufacturing", "Profit (accounting)", "Supply chain", "Peddler", "Bricks and mortar store", "Online shopping", "Corporate strategy", "Market (economics)", "Product (business)", "Customer service", "Wholesale", "Vending machine", "Consumers%27 co-operative", "Department store", "E-commerce", "Retail apocalypse", "Marketing mix", "Product management", "Site selection", "Online marketing platform", "Pricing strategies", "Value-based pricing", "Relationship-based pricing", "Everyday low price", "Product bundling", "Psychological pricing", "Money back guarantee", "Buy one, get one free", "Servicescape", "Marketplace", "Retail chains", "Services marketing", "Disposable product", "Clothing", "Fabrics", "Footwear", "Toiletries", "Cosmetics", "Medicines", "Stationery", "Grocery store", "Supermarkets", "Hypermarkets", "Convenience stores", "Automobiles", "Home appliance", "Electronics", "Furniture", "Sporting goods", "Lumber", "Contemporary art galleries", "Bookstores", "Handicrafts", "Music store", "Gift shops", "Barriers to entry", "National Retail Federation", "Consumer spending", "Gross domestic product", "Food service"

## Appendix 12 Final Curated List of Transportation Wikipedia Articles

"Transportation", "Traffic", "Mode of transport", "Aviation", "Land transport", "Rail transport", "Road transport", "Ship transport", "Cable transport", "Pipeline transport", "Space transport", "Infrastructure", "Vehicle", "Road", "Railway", "Airway (aviation)", "Waterway", "Airport", "Train station", "Bus station", "Warehouse", "Fuel station", "Seaport", "Means of transport", "Wagons", "Automobiles", "Bicycles", "Buses", "Train", "Truck", "Helicopter", "Watercraft", "Spacecraft", "Fixed-wing aircraft", "Public transport", "Private transport", "Air pollution", "Land use", "Traffic flow", "Urban sprawl", "Sustainable transport", "Bicycle", "Rotorcraft", "Gyroplane", "Airliner", "Lift (soaring)", "Landing", "Rail tracks#Railway rail", "Railroad tie", "Rail gauge", "Monorail", "Maglev (transport)", "Locomotive", "Steam locomotive", "Diesel locomotive", "Electric locomotive", "Railway electrification system", "Inter-city rail", "High-speed rail", "Regional rail", "Commuter rail", "Tram", "Rapid transit", "Freight trains", "Box car", "Container train", "Route number", "Trail", "Bus", "Motorcycles", "Pedestrians", "Noise pollution", "Last mile (transportation)", "Barge", "Boat", "Ship", "Sailboat", "Hull (watercraft)", "Steam ships", "Submarine", "Sea transport", "Commercial vessel", "Shipping", "Short sea shipping", "Ferry", "Aerial tramway", "Sea lane", "Airport rail link", "Parking lot", "Transshipment", "Transport finance", "Driving", "People mover", "Passenger", "Flag carrier", "Railway company", "Commuting", "Business travel", "Intermodal passenger transport", "Transport hub", "Railway station", "Demand-responsive transport", "International travel", "Cargo airline", "Transport sustainability", "Transport forecasting", "Transport economics", "Transport engineering", "Trip generation", "Trip distribution", "Mode choice", "Route assignment", "Transport electrification", "Electric car", "Airplane", "Railroads", "Roads", "Vehicles", "Trains", "Public conveyance", "Intersection (road)", "International Regulations for Preventing Collisions at Sea", "Lane", "Junction (traffic)", "Interchange (road)", "Traffic signal", "Traffic cone", "Traffic sign", "Motor vehicle", "Car", "Moped", "Speed limit", "Travel safety", "Road construction", "Car accident", "Debris in the roadway", "Traffic wave", "Gridlock", "Network traffic", "Air traffic", "Driving etiquette", "Vienna Convention on Road Traffic", "Traffic light", "Road traffic control", "Highway Code", "Traffic code", "Uniform Vehicle Code", "Traffic law", "Stop sign", "Vienna Convention on Road Signs and Signals", "Road traffic control device", "Road marking", "Turn signal", "Roundabout", "Priority to the right", "Traffic circle", "Boulevard rule", "All-way stop", "Protected intersection", "Bicycle-friendly", "Pedestrian crossings", "Jaywalking", "Green wave", "Divided highway", "Overtaking", "Left- and right-hand traffic#Left-hand traffic", "Left- and right-hand traffic#Right-hand traffic", "Overtaking#Rules of overtaking", "Road rage", "California Vehicle Code", "Lane splitting", "Grade separation#Roads", "Grade separation", "Underpass", "Interstate highways", "German Autobahn", "Expressways of

China", "Road works", "Traffic jam", "Traffic engineering (transportation)", "Level of service (transportation)", "Three-phase traffic theory", "Rush hour", "Drive time", "Public transportation", "License plate", "High Occupancy Vehicle Lane", "Emergency service", "Fire apparatus", "Winter service vehicle", "Contraflow lane reversal", "Intelligent transportation system", "Floating car data", "Integration of traffic data with navigation systems", "Code of Federal Regulations"

### **Appendix 13 Final Curated List of Healthcare Wikipedia Articles**

"Healthcare", "Health", "Preventive healthcare", "Diagnosis", "COVID-19", "Coronavirus", "SARS-CoV-2", "COVID-19 pandemic", "COVID-19 vaccine", "Long COVID", "Therapy", "Cure", "Disease", "Illness", "Injury", "Disability", "Health professional", "Allied health professions", "Medicine", "Dentistry", "Pharmacy", "Midwifery", "Nursing", "Optometry", "Audiology", "Psychology", "Occupational therapy", "Physical therapy", "Athletic training", "Health profession", "Primary care", "Tertiary care", "Public health", "Health policy", "Health literacy", "Health system", "World Health Organization", "Health Human Resources", "Health facilities", "Physical health", "Mental health", "Well-being", "Eradication of infectious diseases", "Smallpox", "Paraprofessional", "Physiotherapy", "Allied health", "Community health worker", "Unlicensed assistive personnel", "Physical medicine and rehabilitation", "Public health care", "Private healthcare in the United Kingdom", "Health professionals", "Patients", "Health care system", "Primary care physician", "General practitioner", "Family medicine", "Physiotherapist", "Physician assistant", "Nurse practitioner", "Pharmacist", "Nurse", "Patient", "Referral (medicine)", "Urgent care", "Multimorbidity", "Transitional care#continuity", "Preventive medicine", "Health education", "International Classification of Primary Care", "Hypertension", "Diabetes mellitus", "Asthma", "Chronic obstructive pulmonary disease", "Major depressive disorder", "Anxiety disorder", "Back pain", "Osteoarthritis", "Thyroid disease", "Maternal health", "Family planning", "Vaccination", "National Health Interview Survey", "Direct primary care", "Concierge medicine", "Population aging", "Non-communicable disease", "Primary health care", "Acute care", "Hospital", "Emergency department", "Childbirth", "Intensive care medicine", "Medical imaging", "Psychiatrists", "Clinical psychology", "Occupational therapists", "Dental specialties", "Physician", "Health insurance", "Medical specialty", "Medical specialist", "National health insurance", "Respiratory therapists", "Occupational therapist", "Speech and language pathology", "Dietitians", "Inpatient", "Tertiary referral hospital", "Cancer", "Neurosurgery", "Cardiac surgery", "Plastic surgery", "Burn", "Neonatology", "Clinical research", "Surgery", "Self-care", "Home care", "Long-term care", "Assisted living", "Substance use disorder", "Prosthesis", "Orthotics", "Wheelchair", "Health care ratings", "Evaluation", "Healthcare industry", "Biotechnology", "Biopharmaceutical", "Medical model", "Medical diagnosis", "Biomedical research", "Pharmaceutical research", "Evidence-based medicine", "Evidence-based practice", "Health services research", "Social model of disability", "Health care systems", "Health care industry", "Life expectancy", "Health system#International comparisons", "Health administration", "Health department", "Health professional requisites"

### **Appendix 14 Curated Keywords Cultural/Social Activity**

"accompaniment", "activities", "activity", "actors", "acts", "album", "alto", "amusement", "animal", "animals", "anthropologists", "anthropology", "archaeological", "archaeology", "art history", "art", "artifacts", "arts", "asian", "audience", "audio", "auditorium", "band", "bands", "banquet", "banquets", "baroque", "bass", "behavior", "bourdieu", "brass", "camera", "camping", "cantatas", "cappella groups", "cappella music", "cappella", "cello", "century", "ceremonies", "ceremony", "change", "choir", "choirs", "choral music", "choral", "chorus", "christian", "church", "churches", "cinema", "circus", "circuses", "clarinet", "class", "classical music", "classical", "college", "collegiate cappella", "comedy", "community", "competitions", "composers", "concert band", "concert", "concerts", "conductor", "conflict", "consists", "continental", "cooking", "corps", "court", "courts", "cultural capital", "cultural diversity", "cultural heritage", "cultural relativism", "cultural", "culture", "cultures", "cycling", "dance", "dancers", "dances", "distance", "doubling", "drama", "drum", "education", "effects", "ensemble", "ensembles", "entertainment", "episodes", "fair", "fairs", "fan", "fandom", "fans", "fashion", "feast", "festival", "festivals", "fiction", "film festival", "film festivals", "film", "films", "folds", "folk music", "folk", "folklore", "food", "friendships", "fun", "game", "games", "greek", "group", "groups", "guests", "guitar", "habitus", "hairstyles", "hall", "halls", "harmony", "having fun", "heritage", "history", "home cinema", "home", "horn", "human", "individual", "individuals", "instrument", "instrumental", "instruments", "jazz", "leisure", "lgbt", "literature", "magic", "magician", "magicians", "maori", "marx", "masques", "media", "members", "modern", "motets", "multiple aisle", "museology", "museum", "museums", "music ensembles", "music festivals", "music hall", "music", "musical instruments", "musical", "musicians", "myth", "myths", "national", "network", "norm", "norms", "opera", "operas", "operetta", "oral literature", "orchestra", "orchestras", "organ", "palace", "parade", "parties", "parts", "party", "percussion", "perform", "performance", "performed", "performers", "philosophy", "photography", "piano", "play", "player", "players", "playing", "plays", "political", "polyphony", "pop", "popular"

culture", "popular music", "popular", "quartet", "quintet", "race", "races", "racing", "recreation", "recreational", "religion", "religions", "religious", "remix", "renaissance", "restoration", "ritual", "rock", "role", "roles", "roman", "royal", "running", "sacred", "satire", "saxophone", "school", "schools", "science", "scientific", "season", "series", "sing", "singers", "singing", "social class", "social innovation", "social stratification", "social", "socialization", "societies", "society", "sociology", "song", "songs", "soprano", "sound", "south asian", "spectacle", "spectator", "sport", "sports", "sprint finishes", "sprint", "stage", "stratification", "string quartet", "string", "structure", "style", "styles", "sung", "symphony orchestra", "symphony", "tactic", "television", "theater", "theaters", "theatre", "theatres", "theory", "time", "tournament", "tournaments", "tradition", "traditional", "tragedy", "unesco", "university", "venue", "venues", "video game", "video games", "video", "viola", "violins", "vocal group", "vocal", "voice", "voices", "western", "woodwind", "working class", "worship", "writing", "youth", "hobby", "hobbies", # Attractions New York 383 Madison Avenue, "40 Wall Street", "9/11 Tribute Museum", "92nd Street Y", "Abe Lebewohl Park", "Abingdon Square Park", "Adventurer's Park", "African Burial Ground National Monument", "Alben Square", "Albert Capsouto Park", "Alice Austen House", "Alice Tully Hall", "Alley Park", "Alley Pond Park", "American Folk Art Museum", "American International Building", "American Museum of Immigration", "American Museum of Natural History", "America's Response Monument", "Amundsen Circle", "Andrew Heiskell Braille and Talking Book Library", "Anthony Catanzaro Square", "Apollo Theater", "Aqueduct Walk", "Arthur W. Diamond Law Library", "Ascenzi Square", "Asphalt Green", "Asser Levy Recreation Center", "Astoria Park", "Avery Architectural and Fine Arts Library", "Baisley Pond Park", "Bank of America Tower", "Bargemusic", "Barnum's American Museum", "Barretto Point Park", "Bartow-Pell Mansion", "Battery Park", "Battery Park City Ferry Terminal", "Bayswater Point State Park", "Beech tree", "Bellevue South Park", "Belvedere Castle", "Bennett Park", "Bensonhurst Park", "Betsy Head Park", "Billie Holiday Theatre", "Binghamton University", "Blackwell Island Light", "Bloomberg Tower", "Bloomingdale Park", "Blue Heron Park", "Bocchino-Dente Memorial Plaza", "Boone Park", "Bowery Ballroom", "Bowling Green", "Bowne Park", "Bridge Park", "Brill Building", "Broadway theatre", "Bronx Library Center", "Bronx Museum of the Arts", "Bronx Park", "Bronx Skate Park", "Bronx Zoo", "Brookfield Place", "Brooklyn Academy of Music", "Brooklyn Botanic Garden", "Brooklyn Bridge", "Brooklyn Bridge Park", "Brooklyn Children's Museum", "Brooklyn Heights Promenade", "Brooklyn Museum", "Brooklyn Navy Yard", "Brooklyn Public Library", "Brooklyn Symphony Orchestra", "Brooklyn-Queens Greenway", "Brower Park", "Bryant Park", "Buono Beach", "Burton Arms Apartments", "Bush Terminal Park", "Bushwick Inlet Park", "Butler Library", "C.V. Starr East Asian Library", "Cadman Plaza", "Calvert Vaux Park", "campus of New York University", "Captain Patrick J. Brown Walk", "Captain Tilly Park", "Carl Schurz Park", "Carnegie Hall", "Carnegie Hall Tower", "Carroll Park", "Castle Clinton", "Cathedral of St. John the Divine", "Catholic War Veterans Triangle", "Central Park", "Central Park Zoo", "Charlie's Place", "Chelsea Art Museum", "Chelsea Park", "Chelsea Piers", "Chief Charles A. Joshua Plaza", "Children's Museum of the Arts", "Choco-Story New York", "Chris Postiglione Triangle", "Chrysler Building", "Citi Field", "Citigroup Center", "City Hall Park", "City Parks Foundation", "CitySpire Center", "Claremont Park", "Clay Pit Ponds State Park Preserve", "Cleopatra's Needle", "Climate Museum", "Clinton Hall", "Cloisters, The", "Clove Lakes Park", "Cobble Hill Park", "Collect Pond", "College Point Fields", "College Point Little League Building", "Columbia University", "Columbus Circle", "Columbus Park", "Commodore Barry Park", "Con Edison Energy Museum", "Condé Nast Building", "Concrete Plant Park", "Conference House Park", "Conference House", "Conservatory Garden", "Coney Island Creek Park", "Coney Island Cyclone", "Cooper Park", "Cooper-Hewitt, National Design Museum", "Corlears Hook", "Craftsman Music Center", "Crocheron Park", "Crotona Park", "Cunningham Park", "Dag Hammarskjold Library", "Dag Hammarskjold Plaza", "Damrosch Park", "Dante Park", "David Geffen Hall", "David H. Koch Theater", "DeSalvio Playground", "DeWitt Clinton Park", "Diana Ross Playground", "Discovery Times Square Exposition", "Domino Park", "Doughboy Park", "Douglaston Park", "Downtown Athletic Club", "Drumgoole Plaza", "Duane Park", "Dyckman House", "Dyker Beach Park", "Dyker Beach Park and Golf Course", "East Coast Memorial", "East River Esplanade", "East River Greenway", "East River Park", "East River State Park", "Eastern Parkway", "Edgar Allan Poe Cottage", "Eleanor Roosevelt Monument", "Elizabeth H. Berger Plaza", "Ellis Island", "Elmer Holmes Bobst Library", "Elmhurst Park", "Empire State Building", "Estella Diggs Park", "Far Rockaway Skate Park", "Father Demo Square", "Federal Hall", "Feehan Triangle", "Ferry Point Park", "Film Forum", "Film Society of Lincoln Center", "Firemen's Memorial", "Fisher Landau Center", "Flatiron Building", "Flushing Fields", "Flushing Meadows-Corona Park", "Floyd Bennett Field", "Foch Sitting Area", "Foley Square", "Forbes Galleries", "Fordham University", "Forest Park", "Fort George Amusement Park", "Fort Greene Park", "Fort Hill Park", "Fort Schuyler, Bronx", "Fort Tilden", "Fort Totten", "Fort Tryon Park", "Fort Wadsworth", "Fort Washington Park", "Franklin D. Roosevelt Four Freedoms Park", "Freshkills Park", "Frick Art Reference Library", "Frick Collection", "Fulton Park", "GE Building", "Gantry Plaza State Park", "Gateway National Recreation Area", "General Grant National Memorial", "George Gustav Heye Center", "George Washington Bridge", "Givans Creek Woods", "Golconda Skate Park", "Gorman Park", "Gottesman Libraries", "Governors Island", "Governors Island National Monument", "Gracie Mansion", "Gramercy Park", "Grand Army Plaza", "Grand Central Terminal", "Graniteville Quarry Park", "Great Kills Park", "Great Lawn and Turtle Pond", "Green-Wood Cemetery", "Greenacre Park", "Guggenheim Museum SoHo", "Hall of Fame for Great Americans",

"Hamilton Fish Park", "Hamilton Grange National Memorial", "Hammerstein Ballroom", "Hanover Square", "Harlem River Park", "Hart Island", "Harvard Club of New York City", "Heckscher Playground", "Hell's Kitchen Park", "Hendrick I. Lott House", "Henry Hudson Park", "Herald Square", "Herbert Von King Park", "High Line", "Highbridge Park", "Highland Park", "Hillcrest Veterans Square", "Historic Richmond Town", "Hoffman Island", "Holland Tunnel", "Hudson Boulevard", "Hudson Park and Boulevard", "Hudson River Park", "Hunter Island", "Hunts Point Riverside Park", "Ilka Tanya Payán Park", "Imagination Playground", "Imagination Playground at Burling Slip", "Ingram Woods", "International Center of Photography", "International Freedom Center", "Intrepid Sea, Air & Space Museum", "Inwood Hill Park", "Irish Hunger Memorial", "Isham Park", "Isle of Meadows", "J.J. Byrne Park", "J. Hood Wright Park", "Jacob Riis Park", "Jackie Robinson Park", "Jackson Square Park", "Jamaica Bay", "Jamaica Bay Park", "Jamaica Bay Wildlife Refuge", "Jazz at Lincoln Center", "Jerome Park", "Jewish Museum", "Joe Sabba Park", "John J. Carty Park", "John Jay Park", "John Paul Jones Park", "Joseph Rodman Drake Park", "Joyce Kilmer Park", "Judge Moses Weinstein Playground", "Juilliard School", "Julio Carballo Fields", "Juniper Valley Park", "Kaiser Park", "Kaufman Music Center", "KeySpan Park", "King Manor", "Kingsland Homestead", "Kissena Park", "Kohlreiter Square", "Kootyko Triangle", "Kurdish Heritage Foundation of America", "La MaMa Experimental Theatre Club", "Last Chance Pond Park", "Lefferts Historic House", "Leif Ericson Park", "Lemon Creek", "Leon S. Kaiser Playground", "Lewis H. Latimer House", "Libra Triangle", "Liberty Island", "Liberty Park", "Lincoln Center for the Performing Arts", "Lincoln Tunnel", "Linden Park", "Lipstick Building", "Little Island at Pier 55", "Little Red Lighthouse", "LoCicero Triangle", "Long Pond Park", "Louis Cuvillier Park", "Lower East Side Tenement Museum", "Lynch Triangle", "Lyons Pool Recreation Center", "MacArthur Playground", "Macombs Dam Park", "Macri Triangle", "Macy's Herald Square", "Madison Square", "Madison Square Garden", "Mafra Park", "Magenta Playground", "Manhattan Bridge", "Manhattan Municipal Building", "Manhattan School of Music", "Manhattan Waterfront Greenway", "Mannes College of Music", "Marcus Garvey Park", "Maria Hernandez Park", "Marine Park", "Marsha P. Johnson State Park", "Martinez Playground", "McCarren Park", "McGolrick Park", "McKenna Square", "Merchant's House Museum", "Met Breuer", "MetLife Building", "Metropolitan Fireproof Warehouse", "Metropolitan Museum of Art", "Metropolitan Opera", "Midland Beach", "Midtown Manhattan", "Mill Pond Park", "Mill Rock", "Mill Rock Park", "Millennium Park", "Miller Field", "Mitchel Square Park", "Monsignor McGolrick Park", "Montefiore Square", "Morbid Anatomy Museum", "Morgan Library & Museum", "Morningside Park", "Morris-Jumel Mansion", "Mosholu Parkway", "Mount Loretto Unique Area", "Mount Prospect Park", "Mullaly Park", "Municipal Asphalt Plant", "Muscota Marsh", "Museum of American Finance", "Museum of Arts and Design", "Museum of Biblical Art", "Museum of Chinese in America", "Museum of Comic and Cartoon Art", "Museum of Food and Drink", "Museum of Jewish Heritage", "Museum of Modern Art", "Museum of Primitive Art", "Museum of the City of New York", "National Museum of Catholic Art and History", "National Museum of the American Indian", "National September 11 Memorial & Museum", "New Dorp Beach", "New Jersey", "New York Aquarium", "New York Botanical Garden", "New York Chinese Scholar's Garden", "New York City Ballet", "New York City Center", "New York City Hall", "New York City Opera", "New York Evening Post Building", "New York Hall of Science", "New York Jazz Museum", "New York Life Building", "New York Philharmonic", "New York Pops", "New York Public Library", "New York Public Library Main Branch", "New York Public Library for the Performing Arts", "New York Society Library", "New York Stock Exchange", "New York Tattoo Museum", "New York University", "Noguchi Museum", "North and South Brother Islands", "North Woods and North Meadow", "O'Donohue Park", "Ocean Parkway", "Odd Fellows Hall", "Old Stone House", "Onassis Cultural Center", "One Chase Manhattan Plaza", "One Room Schoolhouse Park", "One World Trade Center", "One Worldwide Plaza", "Open Road Park", "Orchard Beach", "Owl's Head Park", "Paley Center for Media", "Paley Park", "Park of the Americas", "Pelham Bay Park", "Pelham Parkway", "Pennsylvania Station", "Peretz Square", "Peter Detmold Park", "Petrosino Square", "Pier 11/Wall Street", "Pier 42", "Plaza Hotel", "Plaza Lafayette", "Playground 52", "Playground Seventy Five", "Poets House", "Prall's Island", "Printer's Park", "Prison Ship Martyrs' Monument", "Proctor-Hopson Circle", "Prospect Park", "Prospect Park Zoo", "Public Theater", "Pugsley Creek Park", "Queen Elizabeth II September 11th Garden", "Queens Botanical Garden", "Queens County Farm Museum", "Queens Museum of Art", "Queens Public Library", "Queens Zoo", "Queensbridge Park", "Rachel Carson Playground", "Racquet and Tennis Club", "Radio City Music Hall", "Ralph Bunche Park", "Randalls and Wards Islands", "Raoul Wallenberg Forest", "Rare Book & Manuscript Library", "Red Hook Park", "Remsen Cemetery", "Riegelmann Boardwalk", "Ripley's Believe It or Not!", "Riverbank State Park", "Riverdale Park", "Riverside Park", "Robert Moses Playground", "Roberto Clemente State Park", "Rock and Roll Hall of Fame", "Rockaway Beach and Boardwalk", "Rockaway Community Park", "Rockefeller Center", "Rodman's Neck", "Rose Center for Earth and Space", "Roundabout Theatre Company", "Roy Wilkins Park", "Rucker Park", "Rufus King Park", "Russell D. Ramsey Triangle", "Sailors' Snug Harbor", "Sakura Park", "Samuel J. Friedman Theatre", "Sara Delano Roosevelt Park", "Saratoga Park", "Schomburg Center for Research in Black Culture", "Science, Industry and Business Library", "Seagram Building", "Seguine Mansion", "Septuagesimo Uno", "Seton Falls Park", "Seth Low Playground", "Seward Park", "Shea Stadium", "Sheridan Square", "Sherman Square", "Shevchenko Scientific Society", "Shirley Chisholm State Park", "Shooters Island", "Shore Boulevard Mall", "Silver Lake", "Silvercup Studios", "Skirball Center for the

Performing Arts", "Snug Harbor Cultural Center", "Society of Illustrators", "Socrates Sculpture Park", "Solomon R. Guggenheim Museum", "Sony Wonder Technology Lab", "Soundview Park", "South Beach Boardwalk", "South Beach–Franklin Delano Roosevelt Boardwalk", "South Street Seaport", "Southpoint Park", "Sphere, The", "Sports Museum of America", "Spring Creek Park", "Spring Street Park", "Springfield Park", "St. James Park", "St. John's Park", "St. Mary's Park", "St. Nicholas Park", "St. Patrick's Cathedral", "St. Vartan Park", "Starlight Park", "Staten Island", "Staten Island Botanical Garden", "Staten Island Greenbelt", "Staten Island Zoo", "Statue of Liberty", "Statue of Liberty National Monument", "Steeplechase Park", "Stonewall National Monument", "Straus Park", "Strawberry Fields (memorial)", "Stuyvesant Cove Park", "Stuyvesant Square", "Sunnyside Gardens", "Sunset Park", "Swedish Cottage Marionette Theatre", "Swinburne Island", "Symphony Space", "Tammany Hall", "Tappen Park", "Taqwa Community Farm", "Tarr Family Playground", "Teardrop Park", "Theodore Roosevelt Birthplace National Historic Site", "Theodore Roosevelt Park", "Thomas Jefferson Park", "Thomas Paine Park", "Times Square", "Tompkins Square Park", "Town Hall", "Travers Park", "Travis Triangle", "Tremont Park", "Triangle 54", "Tribute in Light", "Tribute Park", "Trinity Church", "Trinity Churchyard", "Trump World Tower", "Trygve Lie Plaza", "Turtle Playground", "Udall's Park Preserve", "Udalls Cove", "Underbridge Dog Run", "Union Square", "United Nations Headquarters", "University Plaza", "University Woods", "Valentine-Varian House", "Van Cortlandt House", "Van Cortlandt Park", "Verdi Square", "Vesuvio Playground", "Vietnam Veterans Plaza", "Vincent F. Albano Jr. Playground", "Vinmont Veteran Park", "Vivian Beaumont Theater", "Vleigh Playground", "WNYC Transmitter Park", "Wagner Park", "Waldorf-Astoria Hotel", "Walt Whitman Park", "Walter J. Wetzel Triangle", "Washington Market Park", "Washington Square Park", "Wave Hill", "Wayback Machine", "Weeksville Heritage Center", "West Harlem Piers", "West Side Community Garden", "Whitney Museum of American Art", "William A. Harris Garden", "William T. Davis Wildlife Refuge", "Williamsburg Art & Historical Center", "Williamsburg Bridge", "Williamsbridge Oval", "Willowbrook Park", "Winston Churchill Square", "Wolfe's Pond Park", "Woodlawn Cemetery", "Woolworth Building", "World Trade Center", "Wyckoff-Bennett Homestead", "Yankee Stadium", "Zion Triangle", "Zuccotti Park"

#### **Appendix 15 Curated Keywords Transportation**

"aerial", "air traffic", "aircraft", "airlines", "airplanes", "airport", "airports", "airspace", "airways", "autogyro", "autogyros", "automobile", "aviation", "barge", "bicycle", "bicycles", "boat", "boats", "bus", "buses", "business travel", "car", "cargo", "carriageway", "cars", "commuter rail", "commuter", "convention road", "crossing", "destination", "diesel", "driver", "drivers", "driving", "electric locomotives", "engine", "engines", "expressway", "expressways", "ferries", "ferry", "fixed wing", "flight", "flights", "flying", "freight", "fuel", "fuselage", "glider", "gliders", "helicopter", "helicopters", "high speed", "highway", "infrastructure", "inter city", "interchange", "intersection", "intersections", "jet", "junction", "kite", "km", "landing", "lane", "lanes", "lift", "light rail", "locomotive", "locomotives", "metro", "mile", "monorail", "monorails", "motor", "motorcycles", "motors", "mph", "navigation", "network traffic", "overtaking", "parking", "passenger", "passengers", "pedestrian", "pedestrians", "pilot", "pilots", "plane", "pollution", "port", "ports", "propeller", "public transport", "public transportation", "rail", "railroad", "railroads", "rails", "railway", "railways", "rapid transit", "rapid", "road signs", "road traffic", "road", "roads", "rotor", "rotorcraft", "rotors", "roundabout", "roundabouts", "route", "routes", "ship", "ships", "signal", "signals", "spacecraft", "speed rail", "speed", "stations", "steam locomotives", "stop sign", "stop signs", "submarine", "submarines", "subway", "tail rotor", "terminal", "track", "tracks", "traffic control", "traffic signs", "traffic", "trains", "tram", "tramway", "transit", "transport", "transportation", "transshipment", "travel time", "travel times", "travel", "trip", "trips", "underpasses", "vehicle", "vehicles", "vessels", "wagons", "watercraft", "waterway", "wing aircraft", "driving"

#### **Appendix 16 Curated Keywords Retail Activity**

"ad blocking", "ad server", "ad space", "ad", "ads", "advertisements", "advertisers", "advertising", "amazon", "appliance", "appliances", "automated customer", "banking", "banner ads", "banners", "based pricing", "booksellers", "bookstore", "bookstores", "brand", "bulk", "bundle", "bundles", "bundling", "business users", "business", "businesses", "buy free", "buy", "buyers", "capita", "care products", "cents", "chain", "chains", "cloth", "clothing", "commerce", "commercial", "commercials", "compensation", "competitive", "consumer protection", "consumer spending", "consumer", "consumers", "consumption", "convenience store", "convenience stores", "convenience", "cookies", "cosmetic", "cosmetics", "cost", "costs", "customer service", "customer support", "customer", "customers employees", "customers financial", "customers", "delivered", "delivery", "demand", "department store", "department stores", "display ads", "display advertising", "disposable", "distribution", "economic", "economy", "edlp", "electric", "electronic commerce", "equipment", "exchange", "fabric", "financial institutions", "footwear", "fraud", "furniture", "gdp", "gift", "gni", "good", "goods services", "goods", "grocery stores", "grocery", "growth", "guarantee", "guarantees", "half price", "handicraft", "handicrafts", "hypermarket", "hypermarkets", "industry", "institutional

customers", "institutional", "institutions", "investment", "items", "letterpress", "local", "lumber", "manufacturing", "market", "marketing mix", "marketing", "markets", "materials", "merchandise", "mobile advertising", "money guarantee", "monger", "net", "nrf", "offer", "online ads", "online advertising", "online shopping", "order", "pay", "payment", "peddler", "peddlers", "peddling", "plastic", "price", "prices", "pricing", "product", "production", "products services", "products", "profit", "promotion", "ps", "psychological pricing", "purchase", "purchases", "rate", "rates", "rbp", "retail", "retailers", "revenue", "sale", "sales", "salesman", "sell", "selling", "serve", "service consumer", "service delivery", "service environment", "service experience", "service quality", "service", "services", "servicescape", "servicescapes", "shoes", "shop", "shoppers", "shopping", "shops", "single use", "software", "spam", "spend", "spending", "sponsored", "stationers", "store", "stores", "supermarket", "supermarkets", "supplies", "supply chain", "supply", "tax", "textile", "textiles", "transaction", "transactions", "unit", "value based", "value", "vat", "vending machine", "vending machines", "vending", "wholesale", "wholesalers", "wholesaling", "discount", "discounts"

## **Appendix 17 Curated Keywords Work/Remote Work**

"abolition work", "american worker", "anti trafficking", "anti work", "apprentice", "apprentices", "apprenticeship", "apprenticeships", "basic income", "benefits", "bullshit jobs", "care labor", "care work", "career", "child labor", "child labour", "children work", "collar workers", "collar", "companies", "company", "corporate", "corporations", "coworking spaces", "coworking", "craftsman", "craftsmen", "critique work", "degree", "directors", "division labour", "domestic work", "double burden", "employed", "employee engagement", "employee", "employees", "employer", "employers", "employment", "forced labour", "gdp", "gpi", "graeber", "hour shifts", "hours week", "housework", "human computers", "human trafficking", "income", "isco", "job characteristics", "job guarantee", "job satisfaction", "job", "jobs", "journeymen", "labor force", "labor", "labour force", "labour market", "labour", "limited liability", "management", "managers", "manual labor", "master craftsman", "master craftsmen", "meister", "oecd", "office", "offices", "overwork", "pension", "productivity", "profession", "professional", "professionals", "professions", "protestant work", "qualification", "remote work", "retire", "retirement", "seniority", "sex trafficking", "sex work", "sex workers", "shareholders", "shift", "shifts", "slave trade", "slave", "slavery", "slaves", "social security", "standard working", "tasks", "technician", "tradesman", "traffickers", "trafficking victims", "trafficking", "turnover", "unemployed", "unemployment benefits", "unemployment insurance", "unemployment rate", "unemployment", "unpaid care", "unpaid domestic", "unpaid labor", "unpaid work", "unpaid", "value unpaid", "voluntary", "wage labour", "wage", "wages", "white collar", "work essays", "work ethic", "work family", "work week", "work", "worked", "worker", "workers", "working class", "working hours", "working time", "working", "home office", "wfh", "working from home"

## **Appendix 18 Curated Keywords Healthcare**

"acute", "ageing", "ahps", "allied health", "antidepressant", "antidepressants", "anxiety disorder", "anxiety disorders", "anxiety", "assisted living", "associated", "asthma", "audiology", "behavioral", "biologics", "biotechnology", "birth", "births", "blood pressure", "blood", "burn", "burns", "bypass", "cancer", "cardiac surgery", "cardiac", "care", "cbt", "cells", "childbirth", "chronic", "clinical research", "clinical", "cognitive", "community health", "complications", "contraception", "copd", "coronavirus", "coronaviruses", "corticosteroids", "cosmetic", "cov", "covid 9", "covid", "cure", "cured", "death", "deaths", "december", "dental", "dentistry", "depressed", "depression", "depressive disorder", "depressive", "diabetes", "diagnosis", "diagnostic", "dietitians", "disabilities", "disability", "disabled people", "disabled", "disease", "diseases", "disorder", "disorders", "dr", "drug", "drugs", "dsm", "emergency", "eradication", "family medicine", "family planning", "genome", "health care", "health education", "health insurance", "health literacy", "health professionals", "health professions", "health systems", "health workers", "health", "healthcare", "heart surgery", "heart", "hospital", "hospitals", "human coronavirus", "hypertension", "icpc", "infant", "infants", "infected", "infection", "infections", "inhaled", "injury", "joint", "knee", "labour", "life expectancy", "limb", "long covid", "major depression", "major depressive", "maternal health", "maternal mortality", "maternal", "medical model", "medical", "medication", "medications", "medicine", "mental health", "mental", "midwife", "midwifery", "midwives", "mortality", "multimorbidity", "multiple long", "ncds", "neonatal", "neonatology", "neurosurgery", "newborn", "nhs", "number", "nurse", "nurses", "nursing", "occupational therapist", "occupational therapists", "occupational therapy", "open heart", "optometry", "orthoses", "orthosis", "osteoarthritis", "outbreak", "outpatient", "pain", "pandemic", "pas", "patient", "patients", "pediatrics", "pharmaceutical", "pharmacies", "pharmacist", "pharmacists", "pharmacy", "physical medicine", "physical therapists", "physical therapy", "physical", "physician", "physicians", "physiotherapy", "plastic surgery", "postpartum", "practitioner", "practitioners", "pregnancy", "prevention", "preventive", "primary care", "prosthetic", "psychiatrists", "psychiatry", "psychological", "psychologists", "psychology", "public health", "rehabilitation",

"respiratory", "risk", "rna", "sars cov", "sars", "self care", "severe", "severity", "smallpox", "social anxiety", "specialties", "species", "spinal", "spread", "substance use", "substance", "surgery", "surgical", "symptoms", "syndrome", "tb", "therapist", "therapists", "therapy", "thyroid", "treatment", "uaps", "use disorder", "use disorders", "vaccination", "vaccine", "vaccines", "vaginal", "variant", "variants", "virus", "viruses", "wheelchair", "wheelchairs", "world health", "wuhan"

## References

- ARIFI, D., RESCH, B., SANTILLANA, M., KNOBLAUCH, S., LAUTENBACH, S., JAENISCH, T. & MORALES, I. 2025. How politics affect pandemic forecasting: spatio-temporal early warning capabilities of different geo-social media topics in the context of state-level political leaning. *Frontiers in Public Health*, Volume 13 - 2025.
- ASGARI-CHENAGHLU, M., FEIZI-DERAKHSHI, M.-R., FARZINVASH, L., BALAFAR, M.-A. & MOTAMED, C. 2021. Topic detection and tracking techniques on Twitter: a systematic review. *Complexity*, 2021, 8833084.
- BANDARIN, F., CICIOTTI, E., CREMASCHI, M., MADERA, G., PERULLI, P. & SHENDRIKOVA, D. 2021. After Covid-19: A survey on the prospects for cities. *City, Culture and Society*, 25, 100400.
- BATTY, M., BETTENCOURT, L. M. & KIRLEY, M. 2018. Understanding coupled urban-natural dynamics as the key to sustainability: the example of the galapagos. *Urban Galapagos: Transition to sustainability in complex adaptive systems*. Springer.
- BATTY, M., CLIFTON, J., TYLER, P. & WAN, L. 2022. The post-Covid city. *Cambridge Journal of Regions, Economy and Society*, 15, 447-457.
- BETTENCOURT, L., LOBO, J. & YOUN, H. 2013. The hypothesis of urban scaling: formalization, implications and challenges. *arXiv preprint arXiv:1301.5919*.
- BETTENCOURT, L. M. A. 2013. The Origins of Scaling in Cities. *Science*, 340, 1438-1441.
- BIUK-AGHAI, R. P. & NG, K. K. A method for automated document classification using Wikipedia-derived weighted keywords. 2014 International Conference on Data and Software Engineering (ICODSE), 26-27 Nov. 2014 2014. 1-6.
- BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993–1022.
- CAMPELLO, R. J. G. B., MOULAVI, D. & SANDER, J. Density-Based Clustering Based on Hierarchical Density Estimates. 2013 Berlin, Heidelberg. Springer Berlin Heidelberg, 160-172.
- CAMPELLO, R. J. G. B., MOULAVI, D., ZIMEK, A. & SANDER, J. 2015. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. Data*, 10, Article 5.
- CARPEDM. (n. d.). *GitHub - carpedm20/emoji: emoji terminal output for Python*. [Online]. Available: <https://github.com/carpedm20/emoji/> [Accessed March 2025].
- CHEN, Y., ZHANG, H., LIU, R., YE, Z. & LIN, J. 2019. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, 163, 1-13.
- CONNEAU, A., KHANDELWAL, K., GOYAL, N., CHAUDHARY, V., WENZKE, G., GUZMÁN, F., GRAVE, E., OTT, M., ZETTLEMOYER, L. & STOYANOV, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- CROOKS, A., PFOSE, D., JENKINS, A., CROITORU, A., STEFANIDIS, A., SMITH, D., KARAGIORGOU, S., EFENTAKIS, A. & LAMPRIANIDIS, G. 2015. Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, 29, 720-741.
- CUOMO, A. M. December 21, 2020. Governor Cuomo Announces New York Has Administered 38,000 Doses of COVID-19 Vaccine - Highest Total in the Nation. Available: <https://www.governor.ny.gov/news/governor-cuomo-announces-new-york-has-administered-38000-doses-covid-19-vaccine-highest-total> [Accessed August 2025].
- CUOMO, A. M. February 15, 2021. Governor Cuomo Announces MTA to Partially Restore Overnight Subway Service. Available: <https://www.governor.ny.gov/news/governor-cuomo-announces-mta-partially-restore-overnight-subway-service> [Accessed August 2025].
- CUOMO, A. M. June 15, 2021. Governor Cuomo Announces COVID-19 Restrictions Lifted as 70% of Adult New Yorkers Have Received First Dose of COVID-19 Vaccine. Available: <https://www.governor.ny.gov/news/governor-cuomo-announces-covid-19-restrictions-lifted-70-adult-new-yorkers-have-received-first> [Accessed August 2025].

- CUOMO, A. M. March 20, 2020. Governor Cuomo Signs the 'New York State on PAUSE' Executive Order. Available: <https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order> [Accessed August 2025].
- CUOMO, A. M. October 21, 2020. Governor Cuomo Details COVID-19 Micro-Cluster Metrics. Available: <https://www.governor.ny.gov/news/governor-cuomo-details-covid-19-micro-cluster-metrics> [Accessed August 2025].
- DE ALBUQUERQUE, J. P., HERFORT, B., BRENNING, A. & ZIPF, A. 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29, 667-689.
- DE SABBATA, S., BENNETT, K. & GARDNER, Z. 2023. Towards a study of everyday geographic information: Bringing the everyday into view. *Environment and Planning B: Urban Analytics and City Science*, 23998083231217606.
- DEPARTMENT OF FINANCE (DOF). 2025. *Storefronts Reported Vacant or Not* [Online]. NYC OpenData. Available: [https://data.cityofnewyork.us/City-Government/Storefronts-Reported-Vacant-or-Not/92iy-9c3n/about\\_data](https://data.cityofnewyork.us/City-Government/Storefronts-Reported-Vacant-or-Not/92iy-9c3n/about_data) [Accessed September 2025].
- DEVLIN, J., CHANG, M.-W., LEE, K. & TOUTANOVA, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- EGGER, R. & YU, J. 2022. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7.
- FERRARA, E., VAROL, O., DAVIS, C., MENCZER, F. & FLAMMINI, A. 2016. The rise of social bots. *Communications of the ACM*, 59, 96-104.
- FLORIDA, R., RODRÍGUEZ-POSE, A. & STORPER, M. 2021. Critical Commentary: Cities in a post-COVID world. *Urban Studies*, 60, 1509-1531.
- FOROUHAR, A., CHAPPLE, K., POKHAREL, R. & ALLEN, J. 2025. Transit-driven resilience: Unraveling post-COVID-19 urban recovery dynamics. *Journal of Transport Geography*, 128, 104327.
- FRIAS-MARTINEZ, V., SOTO, V., HOHWALD, H. & FRIAS-MARTINEZ, E. Characterizing urban landscapes using geolocated tweets. 2012 International conference on privacy, security, risk and trust and 2012 international confernece on social computing, 2012. IEEE, 239-248.
- G. ALMATAR, M., ALAZMI, H. S., LI, L. & FOX, E. A. 2020. Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait. *ISPRS International Journal of Geo-Information*, 9, 702.
- GABRILOVICH, E. & MARKOVITCH, S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *IJCAI*, 2007. 1606-1611.
- GALLI, C., DONOS, N. & CALCIOLARI, E. 2024. Performance of 4 Pre-Trained Sentence Transformer Models in the Semantic Query of a Systematic Review Dataset on Peri-Implantitis. *Information*, 15, 68.
- GLAESER, E. L. 2022. Reflections on the post-Covid city. *Cambridge Journal of Regions, Economy and Society*, 15, 747-755.
- GOODCHILD, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, 211-221.
- GROOTENDORST, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- GUAN, X. & CHEN, C. 2014. Using social media data to understand and assess disasters. *Natural Hazards*, 74, 837-850.
- GUO, W., GUPTA, N., POGREBNA, G. & JARVIS, S. Understanding happiness in cities using Twitter: Jobs, children, and transport. 2016 IEEE International Smart Cities Conference (ISC2), 12-15 Sept. 2016 2016a. 1-7.
- GUO, Y., BARNES, S. & JIA, Q. 2016b. Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation. *Tourism Management*, 59, 467-483.
- HARRIS, J. E. 2022. Failure of Concentric Regulatory Zones to Halt the Spread of COVID-19 in South Brooklyn, New York: October-November 2020. *medRxiv*, 2021.11.18.21266493.

- HUANG, J.-H., FLOYD, M. F., TATEOSIAN, L. G. & AARON HIPPI, J. 2022. Exploring public values through Twitter data associated with urban parks pre- and post- COVID-19. *Landscape and Urban Planning*, 227, 104517.
- IGNACCOLO, C., WIBISONO, K., SUTTO, M. P. & PLUNZ, R. A. 2024. Tweeting during the Pandemic in New York City: Unveiling the Evolving Sentiment Landscape of NYC through a Spatiotemporal Analysis of Geolocated Tweets. *Journal of Urban Technology*, 31, 3-28.
- JIANG, B., MA, D., YIN, J. & SANDBERG, M. 2016. Spatial distribution of city tweets and their densities. *Geographical Analysis*, 48, 337-351.
- JIANG, N., CROOKS, A. T., KAVAK, H. & WANG, W. 2023. Leveraging newspapers to understand urban issues: A longitudinal analysis of urban shrinkage in Detroit. *Environment and Planning B: Urban Analytics and City Science*, 23998083231204695.
- JIANG, Y., HUANG, X. & LI, Z. 2021. Spatiotemporal Patterns of Human Mobility and Its Association with Land Use Types during COVID-19 in New York City. *ISPRS International Journal of Geo-Information*, 10, 344.
- JOACHIMS, T. 1997. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc.
- JURDAK, R., ZHAO, K., LIU, J., ABOUJAOUDE, M., CAMERON, M. & NEWTH, D. 2015. Understanding human mobility from Twitter. *PloS one*, 10, e0131469.
- KARAMI, A., LUNDY, M., WEBB, F. & DWIVEDI, Y. K. 2020. Twitter and research: A systematic literature review through text mining. *IEEE access*, 8, 67698-67717.
- KEREDSON. (n. d.). *GitHub - keredson/wordninja: Probabilistically split concatenated words using NLP based on English Wikipedia unigram frequencies* [Online]. Available: <https://github.com/keredson/wordninja> [Accessed March 2025].
- KONTOKOSTA, C. E., FREEMAN, L. & LAI, Y. 2024. Up-and-Coming or Down-and-Out? Social Media Popularity as an Indicator of Neighborhood Change. *Journal of Planning Education and Research*, 44, 662-673.
- KOVACS-GYÖRI, A., RISTEA, A., KOLCSAR, R., RESCH, B., CRIVELLARI, A. & BLASCHKE, T. 2018. Beyond Spatial Proximity—Classifying Parks and Their Visitors in London Based on Spatiotemporal and Sentiment Analysis of Twitter Data. *ISPRS International Journal of Geo-Information* [Online], 7.
- LANSLEY, G. & LONGLEY, P. A. 2016. The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58, 85-96.
- LI, X., XU, M., ZENG, W., TSE, Y. K. & CHAN, H. K. 2023. Exploring customer concerns on service quality under the COVID-19 crisis: A social media analytics study from the retail industry. *Journal of Retailing and Consumer Services*, 70, 103157.
- LI, Z., HUANG, X., YE, X., JIANG, Y., MARTIN, Y., NING, H., HODGSON, M. E. & LI, X. 2021. Measuring global multi-scale place connectivity using geotagged social media data. *Scientific Reports*, 11, 14694.
- LIAO, Y., YEH, S. & GIL, J. 2022. Feasibility of estimating travel demand using geolocations of social media data. *Transportation*, 49, 137-161.
- LIU, Y., SUI, Z., KANG, C. & GAO, Y. 2014. Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. *PLOS ONE*, 9, e86026.
- MCCORRISTON, J., JURGENS, D. & RUTHS, D. Organizations are users too: Characterizing and detecting the presence of organizations on twitter. *Proceedings of the international aaai conference on web and social media*, 2015. 650-653.
- MCINNES, L., HEALY, J. & MELVILLE, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- MEGAHED, N. A. & ABDEL-KADER, R. F. 2022. Smart Cities after COVID-19: Building a conceptual framework through a multidisciplinary perspective. *Scientific African*, 17, e01374.
- MILLER, H. J. & GOODCHILD, M. F. 2015. Data-driven geography. *GeoJournal*, 80, 449-461.

- NAGARKAR, P., KHAN, A., RAIKAR, S. & ZANTYE, A. Twitter Data Mining for Targeted Marketing. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 15-17 July 2020 2020. 44-50.
- NGUYEN, H. L., TSOLAK, D., KARMANN, A., KNAUFF, S. & KÜHNE, S. 2022. Efficient and Reliable Geocoding of German Twitter Data to Enable Spatial Data Linkage to Official Statistics and Other Data Sources. *Frontiers in Sociology*, 7.
- NIU, H. & SILVA, E. A. 2020. Crowdsourced Data Mining for Urban Activity: Review of Data Sources, Applications, and Methods. *Journal of Urban Planning and Development*, 146, 04020007.
- NIU, H. & SILVA, E. A. 2023. Understanding temporal and spatial patterns of urban activities across demographic groups through geotagged social media data. *Computers, Environment and Urban Systems*, 100, 101934.
- NYC DEPARTMENT OF HEALTH AND MENTAL HYGIENE (DOHMH). *NYC Coronavirus (COVID-19) data (2020)* [Online]. Available: <<https://github.com/nychealth/coronavirus-data>> [Accessed August 2025].
- NYC MAYOR'S OFFICE [@NYCMAYORSOFFICE]. 2020. *If you live in a #RedZone, you're closest to the latest COVID-19 outbreaks. You'll see strict containment measures. Help keep our city safe, and learn what zone you live in at* <http://NYC.gov/COVIDZone> [Online]. X. Available: <https://x.com/NYCMayorsOffice/status/1315321260184801280> [Accessed August 2025].
- NYC OFFICE OF TECHNOLOGY & INNOVATION. 2022a. *Ferry Landings* [Online]. ArcGIS Online. Available: [https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/FerryLandings\\_View/FeatureServer](https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/FerryLandings_View/FeatureServer) [Accessed September 2025].
- NYC OFFICE OF TECHNOLOGY & INNOVATION. 2022b. *Railroad 2022* [Online]. ArcGIS Online. Available: [https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/Railroad\\_2022/FeatureServer](https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/Railroad_2022/FeatureServer) [Accessed September 2025].
- NYC OFFICE OF TECHNOLOGY & INNOVATION. 2024a. *Facility Database (DCP)* [Online]. ArcGIS Online. Available: [https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/Park\\_2022/FeatureServer](https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/Park_2022/FeatureServer) [Accessed September 2025].
- NYC OFFICE OF TECHNOLOGY & INNOVATION. 2024b. *Park 2022* [Online]. ArcGIS Online. Available: [https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/Park\\_2022/FeatureServer](https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/Park_2022/FeatureServer) [Accessed September 2025].
- NYC OFFICE OF TECHNOLOGY & INNOVATION. 2025a. *Subway* [Online]. ArcGIS Online. Available: [https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/Subway\\_view/FeatureServer](https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/Subway_view/FeatureServer) [Accessed September 2025].
- NYC OFFICE OF TECHNOLOGY & INNOVATION. 2025b. *Subway Station* [Online]. ArcGIS Online. Available: [https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/SubwayStation\\_view/FeatureServer](https://services6.arcgis.com/yG5s3afENB5iO9fj/arcgis/rest/services/SubwayStation_view/FeatureServer) [Accessed September 2025].
- PATWARDHAN, N., MARRONE, S. & SANSONE, C. 2023. Transformers in the real world: A survey on nlp applications. *Information*, 14, 242.
- QIANG, D. & MCKENZIE, G. 2024. Navigating the post-pandemic urban landscape: Disparities in transportation recovery & regional insights from New York City. *Computers, Environment and Urban Systems*, 110, 102111.
- RAJPUT, A. A., LI, Q., GAO, X. & MOSTAFAVI, A. 2022. Revealing Critical Characteristics of Mobility Patterns in New York City During the Onset of COVID-19 Pandemic. *Frontiers in Built Environment*, Volume 7 - 2021.
- RAMOS, J. 2003. Using TF-IDF to determine word relevance in document queries.
- RAO, F., HAN, S. S. & PAN, R. 2022. Planning for resilient central-city shopping districts in the post-Covid era: an explanatory case study of the Hoddle Grid in Melbourne. *Cambridge Journal of Regions, Economy and Society*, 15, 575-596.

- RAPIDSAL. (n. d.). *GitHub - rapidsai/cuml: cuML - RAPIDS Machine Learning Library* [Online]. Available: <https://github.com/rapidsai/cuml> [Accessed June 2025].
- RATH, S. & CHOW, J. Y. J. 2022. Worldwide city transport typology prediction with sentence-BERT based supervised learning via Wikipedia. *Transportation Research Part C: Emerging Technologies*, 139, 103661.
- REIMERS, N. & GUREVYCH, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- RESCH, B., USLÄNDER, F. & HAVAS, C. 2018. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, 45, 362-376.
- REUSCHKE, D., LONG, J. & BENNETT, N. 2021. Locating creativity in the city using Twitter data. *Environment and Planning B: Urban Analytics and City Science*, 48, 2607-2622.
- RICHARDSON, L. 2024. *Beautiful Soup Documentation* [Online]. Available: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> [Accessed November 2024].
- ROBERTS, H., RESCH, B., SADLER, J., CHAPMAN, L., PETUTSCHNIG, A. & ZIMMER, S. 2018. Investigating the Emotional Responses of Individuals to Urban Green Space Using Twitter Data: A Critical Comparison of Three Different Methods of Sentiment Analysis. *Urban Planning; Vol 3, No 1 (2018): Crowdsourced Data and Social Media in Participatory Urban Planning*.
- SALTON, G. & BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24, 513-523.
- STEIGER, E., RESCH, B. & ZIPF, A. 2016. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, 30, 1694-1716.
- U.S. CENSUS BUREAU. *Decennial Census, DEC Redistricting Data (PL 94-171), Table P1* [Online]. Available: <https://data.census.gov/table/DECENNIALPL2020.P1?g=160XX00US3651000> [Accessed August 2025].
- U.S. CENTERS FOR DISEASE CONTROL AND PREVENTION. *CDC Museum COVID-19 Timeline* [Online]. Available: <https://www.cdc.gov/museum/timeline/covid19.html> [Accessed August 2025].
- VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. & POLOSUKHIN, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- WANG, D., HE, B. Y., GAO, J., CHOW, J. Y. J., OZBAY, K. & IYER, S. 2021. Impact of COVID-19 behavioral inertia on reopening strategies for New York City transit. *International Journal of Transportation Science and Technology*, 10, 197-211.
- WU, Z., ZHU, H., LI, G., CUI, Z., HUANG, H., LI, J., CHEN, E. & XU, G. 2017. An efficient Wikipedia semantic matching approach to text document classification. *Information Sciences*, 393, 15-28.
- XU, W. W., TSHIMULA, J. M., DUBÉ, È., GRAHAM, J. E., GREYSON, D., MACDONALD, N. E. & MEYER, S. B. 2022. Unmasking the Twitter discourses on masks during the COVID-19 pandemic: User cluster-based BERT topic modeling approach. *Jmir Infodemiology*, 2, e41198.
- YANG, C. & LIU, T. 2022. Social Media Data in Urban Design and Landscape Research: A Comprehensive Literature Review. *Land* [Online], 11.
- ZHAO, H., MAILLOUX, B. J., COOK, E. M. & CULLIGAN, P. J. 2023. Change of urban park usage as a response to the COVID-19 global pandemic. *Scientific Reports*, 13, 19324.
- ZHONG, C., MORPHET, R. & YOSHIDA, M. 2023. Twitter mobility dynamics during the COVID-19 pandemic: A case study of London. *PLoS One*, 18, e0284902.